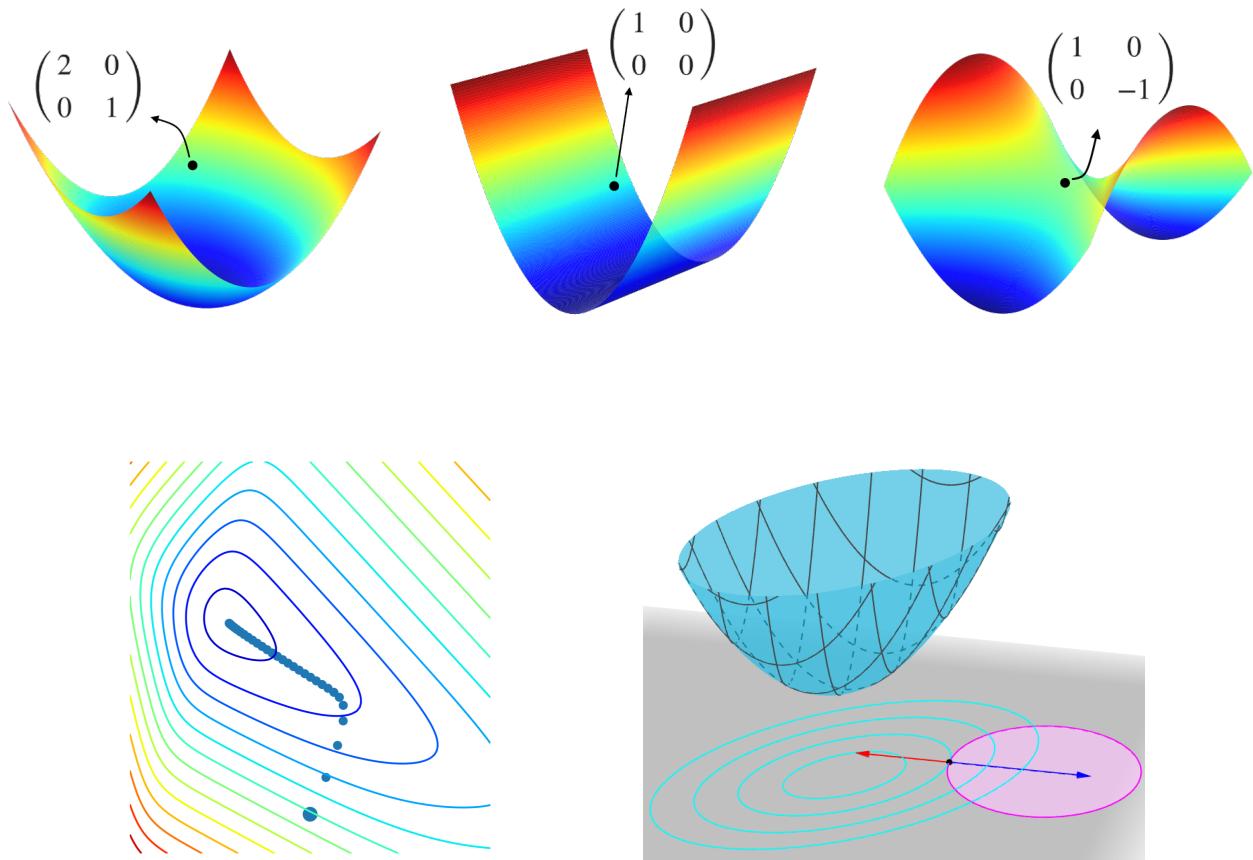

Optimisation

Guillaume Garrigos



Guide de lecture

Prérequis : Notions d'Algèbre Linéaire et de Calcul Différentiel. Les notions dont nous aurons besoin pour ce cours sont réunies dans le Chapitre I, qui sert d'introduction à ce cours. En particulier, il est nécessaire d'avoir une bonne compréhension de ce que sont les matrices (semi-)définies positives, et le gradient et la hessienne d'une fonction à valeurs réelles.

Hors-piste : Les sections dont le titre se termine par une astérisque * sont plus avancées. Elles sont donc, par défaut, hors programme, à moins que le temps nous permette de les traiter en cours. Elles permettent dans tout les cas d'apporter des informations complémentaires, qui je l'espère satisferont les plus curieuses et curieux. C'est le cas des Annexes, qui contiennent les preuves de résultats qui ont été admis pendant le cours, ainsi que des développements un peu plus avancés.

Références

Ces notes de cours ont été rédigées entre 2020-2021, sur la base d'un polycopié d'Olivier Bokanowski, ajourné par Matthieu Bonnivard. Au cas où le contenu de ce polycopié ne vous suffise pas, voici quelques références qui vous permettront d'aller plus loin.

- *Optimisation et analyse convexe : Exercices et problèmes corrigés, avec rappels de cours* par Jean-Baptiste Hiriart-Urruty [7]. L'auteur est un très bon pédagogue et agréable à lire. Comme le suggère le titre, son livre contient de nombreux exercices corrigés. Attention toutefois, son contenu est de difficulté variable, avec des chapitres qui dépassent le cadre de ce cours. Focalisez-vous sur les 3 premiers chapitres (sauf III.2). Je ne peux que vous inviter à lire également la section historique en fin du livre, riche en anecdotes.
- *Objectif Agrégation*, par Vincent Beck, Jérôme Malick et Gabriel Peyré [3]. Voici également un livre que je trouve très bien écrit, certainement un de mes préférés. C'est un livre généraliste (qui couvre analyse et algèbre), mais son premier chapitre donne une vision d'ensemble sur le calcul différentiel et ses applications qui je pense vaut le coup d'œil.
- *Nonlinear Programming*, par Dimitri P. Bertsekas [4]. L'auteur est bon pédagogue, et accompagne ses explications par des dessins et schémas très utiles à la compréhension. Les chapitres 1.1-4 portent sur le contenu des chapitres II et IV. Le chapitre 3, en particulier la partie 3.3, développe en détail le contenu du chapitre V.
- *Introduction à l'analyse numérique matricielle et à l'optimisation*, par Philippe Ciarlet [5]. Un classique, mais qui a un peu vieilli. Le chapitre 1 vous fournira de bons rappels en Algèbre Linéaire. Les chapitres 7.1-4 et 8.1-4 portent sur le contenu du cours, le reste dépasse le cadre du cours.
- *Analyse numérique et optimisation : Une introduction à la modélisation mathématique et à la simulation numérique*, par Grégoire Allaire [1]. Ce livre se focalise sur la résolution des Équations aux Dérivées Partielles, et ses chapitres 9-10 fournissent des exemples intéressants d'application des résultats de ce cours aux EDPs. Attention cependant, l'auteur travaille dans le cadre d'espaces de Hilbert, et sa présentation des résultats diffère du contenu de ce cours et parfois dépasse son cadre.

Table des matières

I Éléments d'Algèbre Linéaire et de Calcul Différentiel	9
I.I Rappels et compléments d'algèbre linéaire	10
I.I.1 La structure euclidienne de \mathbb{R}^N	10
I.I.2 Spectre d'une matrice carrée	12
I.I.3 L'algèbre normée $\mathcal{M}_{M,N}(\mathbb{R})$	14
I.I.4 Matrices symétriques et antisymétriques	15
I.I.5 Matrices semi-définies positives et définies positives	17
I.II Rappels et compléments de calcul différentiel	23
I.II.1 Différentielle	23
I.II.2 Différentielle seconde	26
I.II.3 Fonctions quadratiques	27
II Existence de minimiseurs et conditions d'optimalité	29
II.I Conditions d'optimalité et Principe de Fermat	29
II.I.1 Un peu de vocabulaire	30
II.I.2 Conditions d'Optimalité du 1er ordre	32
II.I.3 Conditions d'Optimalité du 2e ordre	33
II.II Coercivité et existence de minimiseurs	36
II.II.1 Coercivité	36
II.II.2 Existence de minimiseurs	38
II.III Récapitulatif du Chapitre	41
III Optimisation convexe	43
III.I Convexité et globalité des minimiseurs	43
III.I.1 Ensemble convexe	43
III.I.2 Fonction convexe	44
III.I.3 Caractérisation de la convexité pour les fonctions univariées	46
III.I.4 Caractérisation de la convexité pour les fonctions multivariées	48
III.I.5 Convexité et minimiseurs	51
III.II Forte convexité : existence et unicité du minimiseur	52
III.II.1 Fonction fortement convexe	52
III.II.2 Caractérisation de la forte convexité	53

III.II.3 Forte convexité et minimiseurs	54
III.III Récapitulatif du Chapitre	56
IV Algorithmes de minimisation sans contrainte	57
IV.I Méthodes de descente	57
IV.I.1 Algorithmes itératifs	57
IV.I.2 Directions de descente	59
IV.I.3 Méthodes du gradient et de Newton	61
IV.II Conditionnement	64
IV.II.1 Fonctions à gradient Lipschitzien	64
IV.II.2 Conditionnement d'une fonction	67
IV.III Méthode du gradient	69
IV.III.1 La méthode du gradient à pas fixe	69
IV.III.2 Méthode du gradient à pas optimal	75
IV.IV Récapitulatif du Chapitre	79
V Optimisation sous contraintes	81
V.I Introduction : Problèmes classiques	81
V.I.1 Polyèdres	82
V.I.2 Optimisation Linéaire	85
V.I.3 Optimisation Convexe	87
V.II Théorème(s) de Lagrange-KKT	90
V.II.1 Contrainte d'inégalité simple et multiplicateur	90
V.II.2 Condition d'Optimalité de KKT du 1er ordre	94
V.II.3 Condition d'Optimalité de KKT du 2e ordre	103
V.III Algorithmes pour l'optimisation sous contraintes	107
V.III.1 Projection sur un convexe fermé	107
V.III.2 Propriétés avancées de la projection	112
V.III.3 Algorithme du gradient projeté	113
V.III.4 Algorithme de projection alternées *	117
V.III.5 Pour aller plus loin *	120
V.IV Récapitulatif du Chapitre	122
A Annexe : Convexité(s) et Convergence *	123

Chapitre I

Éléments d'Algèbre Linéaire et de Calcul Différentiel

L'optimisation est une discipline qui emprunte beaucoup de notions à l'algèbre linéaire et au calcul différentiel. Voici donc quelques rappels concernant les notions dont vous aurez besoin dans ce cours. Les résultats qui suivent sont admis, bien que pour certains nous reverrons leurs preuves en TD. J'en profite également pour tordre le cou à certaines idées préconçues.

Comment lire ce chapitre? Ceci est essentiellement un chapitre de rappels, bien qu'il puisse contenir des choses que vous n'avez pas vues, ou simplement oubliées. Je vous conseille donc d'en faire une première lecture en diagonale, afin de déterminer si ce qui s'y trouve vous semble familier ou non ; puis, dans un deuxième temps, de travailler les parties qui vous semblent les plus obscures. Vous pourrez par exemple vous tourner vers les exercices qui sont proposés, que vous trouverez également dans la feuille de TD. Ils ne seront pas tous traités en TD, donc n'hésitez pas à en piocher quelques-uns par vous-mêmes.

Notations.

- N, M désigneront toujours des entiers supérieurs ou égaux à 1.
- $\mathcal{L}(E; F)$ désigne l'espace des applications linéaires entre les espaces vectoriels E et F .
- $\mathcal{B}(E, F; G)$ désigne l'espace des applications bilinéaires entre les espaces vectoriels $E \times F$ et G .
- \mathbb{R}_+ (resp. \mathbb{R}_-) est une notation pour $[0, +\infty[$ (resp. $] -\infty, 0]$).

I.I Rappels et compléments d'algèbre linéaire

Dans ce cours, on note $\mathcal{M}_{M,N}(\mathbb{R})$ l'espace vectoriel des matrices à M lignes et N colonnes. Si $M = N$ on écrira simplement $\mathcal{M}_N(\mathbb{R})$. La transposée d'une matrice $A \in \mathcal{M}_{M,N}(\mathbb{R})$ se notera A^\top , ou parfois A^* (il subsiste encore quelques doublons qu'il faut supprimer). Par défaut, les vecteurs de \mathbb{R}^N exprimés dans la base canonique seront considérés comme des éléments de $\mathcal{M}_{N,1}(\mathbb{R})$, c'est-à-dire des vecteurs « colonne ».

I.I.1 La structure euclidienne de \mathbb{R}^N

I.I.1.i) Définitions de base

Le produit scalaire euclidien dans \mathbb{R}^N , noté $\langle \cdot, \cdot \rangle : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}$, est défini par :

$$(\forall x \in \mathbb{R}^N)(\forall y \in \mathbb{R}^N) \quad \langle x, y \rangle := \sum_{i=1}^N x_i y_i.$$

Si on regarde les vecteurs de \mathbb{R}^N comme des vecteurs colonne, on peut également écrire le produit scalaire comme un produit matriciel entre une ligne et une colonne : $\langle x, y \rangle = x^\top y$.

La norme euclidienne de \mathbb{R}^N , notée $\| \cdot \| : \mathbb{R}^N \rightarrow \mathbb{R}_+$, est définie par :

$$(\forall x \in \mathbb{R}^N) \quad \|x\| := \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^N x_i^2}.$$

La distance euclidienne de \mathbb{R}^N , notée $d(\cdot, \cdot) : \mathbb{R}^N \times \mathbb{R}^N \rightarrow \mathbb{R}_+$, est définie par :

$$(\forall x \in \mathbb{R}^N)(\forall y \in \mathbb{R}^N) \quad d(x, y) := \|x - y\| = \sqrt{\sum_{i=1}^N (x_i - y_i)^2}.$$

Voici quelques propriétés utiles pour faire des calculs incluant des produits scalaires et des normes :

Proposition I.1.

- i) (Identité remarquable 1) Pour tous $x, y \in \mathbb{R}^N$, $\|x + y\|^2 = \|x\|^2 + \|y\|^2 + 2\langle x, y \rangle$.
- ii) (Identité remarquable 2) Pour tous $x, y \in \mathbb{R}^N$, $\|x\|^2 - \|y\|^2 = \langle x + y, x - y \rangle$.
- iii) (Inégalité de Cauchy-Schwarz) Pour tous $x, y \in \mathbb{R}^N$, $-\|x\|\|y\| \leq \langle x, y \rangle \leq \|x\|\|y\|$.
- iv) (Règle de l'adjoint) Pour toute matrice $A \in \mathcal{M}_{M,N}(\mathbb{R})$, $x \in \mathbb{R}^N$, $y \in \mathbb{R}^M$, $\langle Ax, y \rangle = \langle x, A^\top y \rangle$.

Remarque I.2. Cette quatrième propriété est souvent méconnue/oubliée par les étudiant(e)s. Elle est pourtant essentielle pour tout les calculs impliquant matrice et produit scalaire. On la retrouvera régulièrement au long de ce cours. Elle permet par exemple d'écrire des choses comme $\|Ax\|^2 = \langle A^\top Ax, x \rangle$.

I.I.1.ii) Orthogonalité

Définition I.3. On dira que deux vecteurs x et y de \mathbb{R}^N sont **ORTHOGONAUX** lorsque $\langle x, y \rangle = 0$.

Remarque I.4. C'est une notion que vous avez rencontré à de multiples reprises, par exemple les bases orthogonales (bases dont les vecteurs sont tous orthogonaux les uns avec les autres).

Définition I.5. Soit $F \subset \mathbb{R}^N$ un sous-espace vectoriel. On dit que $x \in \mathbb{R}^N$ est orthogonal à F s'il est orthogonal avec tous les vecteurs de F . On définit l'**ORTHOGONAL** de F comme étant l'ensemble de tous les vecteurs orthogonaux à F :

$$F^\perp := \{x^* \in \mathbb{R}^N \mid (\forall x \in F) \quad \langle x^*, x \rangle = 0\}.$$

Proposition I.6.

- i) F^\perp est un sous-espace vectoriel de \mathbb{R}^N .
- ii) F et F^\perp sont supplémentaires. En particulier, $\dim F + \dim F^\perp = N$.
- iii) $(F^\perp)^\perp = F$.

Un résultat très utile :

Proposition I.7. Soit $A \in \mathcal{M}_N(\mathbb{R})$. Alors $\text{Ker}(A)^\perp = \text{Im}(A^\top)$ et $\text{Im}(A)^\perp = \text{Ker}(A^\top)$.

I.I.1.iii) Topologie euclidienne dans \mathbb{R}^N .

Quelques définitions :

Définition I.8. Soient $x \in \mathbb{R}^N$ et $r \in]0, +\infty[$. On définit

- La **BOULE OUVERTE** centrée en x , de rayon r , par

$$\mathbb{B}(x, r) := \{y \in \mathbb{R}^N \mid d(x, y) < r\}.$$

- La **BOULE FERMÉE** centrée en x , de rayon r , par

$$\overline{\mathbb{B}}(x, r) := \{y \in \mathbb{R}^N \mid d(x, y) \leq r\}.$$

Définition I.9.

- On dit qu'un ensemble $U \subset \mathbb{R}^N$ est **OUVERT** si

$$(\forall x \in U)(\exists r > 0) \quad \mathbb{B}(x, r) \subset U.$$

- On dit qu'un ensemble $F \subset \mathbb{R}^N$ est **FERMÉ** si son complémentaire $\mathbb{R}^N \setminus F$ est ouvert.

- Étant donné un ensemble $C \subset \mathbb{R}^N$, on définit son **INTÉRIEUR**, que l'on note $\text{int } C$, comme étant l'ensemble

$$\text{int } C := \{x \in C \mid (\exists r > 0) \quad \mathbb{B}(x, r) \subset C\}.$$

Remarque I.10. Par définition, l'intérieur d'un ensemble est le plus petit ouvert inclus dans cet ensemble. Ces définitions impliquent également que la boule ouverte est ouverte, et que la boule fermée est fermée (heureusement!).

I.I.2 Spectre d'une matrice carrée

Définition I.11. On dit que $\lambda \in \mathbb{R}$ est une **VALEUR PROPRE** (réelle) de $A \in \mathcal{M}_N(\mathbb{R})$ s'il existe un vecteur non nul $x \in \mathbb{R}^N$ tel que $Ax = \lambda x$. Autrement dit, si $A - \lambda I$ n'est pas inversible dans $\mathcal{M}_N(\mathbb{R})$. On note $\text{spec}_{\mathbb{R}}(A)$ l'ensemble des valeurs propres de A .

Proposition I.12. Les valeurs propres de $A \in \mathcal{M}_N(\mathbb{R})$ sont les racines réelles du polynôme caractéristique $X \mapsto \det(XI_N - A)$.

Remarque I.13. Une matrice $A \in \mathcal{M}_N(\mathbb{R})$ peut ne posséder aucune valeur propre. Par exemple la matrice

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

dont le polynôme caractéristique est $X^2 + 1$.

Définition I.14. On dit que $\lambda \in \mathbb{C}$ est une **VALEUR SPECTRALE** (ou valeur propre complexe) de $A \in \mathcal{M}_N(\mathbb{R})$ s'il existe un vecteur non nul $x \in \mathbb{C}^N$ tel que $Ax = \lambda x$. Autrement dit, si $A - \lambda I$ n'est pas inversible dans $\mathcal{M}_N(\mathbb{C})$. Le **SPECTRE** de A , noté $\text{spec}(A)$, est l'ensemble des valeurs spectrales de A .

Proposition I.15. Les valeurs spectrales de $A \in \mathcal{M}_N(\mathbb{R})$ sont les racines complexes du polynôme caractéristique $\det(XI_N - A)$.

Corollaire I.16. Pour $A \in \mathcal{M}_N(\mathbb{R})$, $\text{spec}_{\mathbb{R}}(A) = \text{spec}(A) \cap \mathbb{R}$.

Remarque I.17.

- Dans certains cas, toutes les valeurs spectrales sont réelles : $\text{spec}(A) = \text{spec}_{\mathbb{R}}(A)$. On va par exemple voir que c'est le cas pour les matrices symétriques.
- Le spectre n'est jamais vide. C'est une conséquence du fait que tout polynôme réel admet au moins une racine dans \mathbb{C} .

Proposition I.18. Si $A \in \mathcal{M}_N(\mathbb{R})$ est triangulaire, alors

$$\text{spec}(A) = \text{spec}_{\mathbb{R}}(A) = \{A_{11}, \dots, A_{NN}\}.$$

Remarque I.19. Pour les matrices triangulaires, et en particulier pour les matrices diagonales, les valeurs propres se situent donc sur la diagonale. C'est très pratique! Mais c'est malheureusement faux en règle générale. Par exemple, le spectre de $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ est $\{-i, +i\}$, qui ne contient pas $\{0\}$.

Voici un résultat classique sur les valeurs spectrales d'une matrice :

Proposition I.20. Soit $A \in \mathcal{M}_N(\mathbb{R})$, et soient $\lambda_1, \dots, \lambda_N$ les valeurs spectrales de A , comptées avec leur multiplicité algébrique. Alors

- i) $\text{tr}(A) = \sum_{i=1}^N \lambda_i,$
- ii) $\det(A) = \prod_{i=1}^N \lambda_i.$

Remarque I.21. Il est important de prendre en compte la « multiplicité algébrique » ici! Par exemple, considérons l'exemple très simple de la matrice $2I_3$:

$$A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}.$$

Il est clair que $\text{spec}(A) = \{2\}$ (et non pas $(2, 2, 2)$: on parle d'*ensemble*, pas de *uplet*!), c'est à dire qu'il y a une unique valeur spectrale 2. Pour autant on voit bien que $\text{tr}(A) \neq 2$ et $\det(A) \neq 2$. Pour que ce résultat marche, il nous faut prendre en compte la *multiplicité algébrique* de 2. Cette multiplicité est exactement la puissance apparaissant dans le polynôme caractéristique de A , qui est ici $(X - 2)^3$.

Définition I.22. Le **RAYON SPECTRAL** d'une matrice $A \in \mathcal{M}_N(\mathbb{R})$, noté $\rho(A)$, est défini par

$$\rho(A) := \max\{|\lambda| \mid \lambda \in \text{spec}(A)\}.$$

Remarque I.23. Un contre-sens classique est de penser que « le rayon spectral est la plus grande valeur propre ». Ceci est faux, pour de nombreuses raisons :

- Les valeurs propres peuvent ne pas exister. Le rayon spectral porte sur les valeurs spectrales (ou les valeurs propres complexes).
- On ne peut pas parler de « plus grande valeur spectrale » non plus, car \mathbb{C} n'est pas muni d'une relation d'ordre total, contrairement à \mathbb{R} ! On ne peut pas comparer $2i$ et $1+i$ par exemple. Par contre on peut comparer leur *module* 2 et $\sqrt{2}$.

- Même lorsque le spectre est réel, le rayon spectral ne maximise pas les valeurs propres mais leur *valeur absolue*. Par exemple, pour la matrice

$$A = \begin{pmatrix} 1 & 0 \\ 0 & -2 \end{pmatrix}$$

la plus grande valeur propre est 1 (puisque $1 > -2$), mais $\rho(A) = 2$. Cela peut paraître un détail mais cela a son importance !

I.I.3 L'algèbre normée $\mathcal{M}_{M,N}(\mathbb{R})$

Définition I.24. La **NORME D'OPÉRATEUR** d'une matrice $A \in \mathcal{M}_{M,N}(\mathbb{R})$ (on parle aussi de norme subordonnée euclidienne) est définie par :

$$\|A\| := \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

Remarque I.25 (NormeS matricielleS). Il existe de nombreuses façons de munir $\mathcal{M}_{M,N}(\mathbb{R})$ d'une norme. Ceux parmi vous ayant suivi le cours d'Analyse Numérique Matricielle en auront vu une palanquée (les normes d'opérateur ℓ^p/ℓ^q , la norme de Frobenius) et il en existe bien d'autres (citons la très utile norme nucléaire), les plus curieux pourront consulter l'article Wikipédia sur le sujet¹. Néanmoins, dans ce cours nous ferons seulement appel à la norme d'opérateur subordonnée à la norme euclidienne $\|\cdot\|$ mentionnée ci-dessus.

Cette norme d'opérateur $\|\cdot\|$ vérifie deux inégalités très importantes. La première est une conséquence directe de la définition. La seconde est une propriété de sous-multiplicativité, qui fait de $\|\cdot\|$ ce que l'on appelle une norme d'algèbre.

Proposition I.26.

- i) $\|\cdot\|$ est une norme sur $\mathcal{M}_{M,N}(\mathbb{R})$.
- ii) Pour tout $A \in \mathcal{M}_{M,N}(\mathbb{R})$, $x \in \mathbb{R}^N$, $\|Ax\| \leq \|A\| \|x\|$.
- iii) Pour tout $A \in \mathcal{M}_{M,N}(\mathbb{R})$, $B \in \mathcal{M}_{N,P}(\mathbb{R})$, $\|AB\| \leq \|A\| \|B\|$.

Exercice I.27. Soit $A \in \mathcal{M}_N(\mathbb{R})$ telle que $\|A\| < 1$. Montrer que A^k tend vers 0 (la matrice nulle) lorsque $k \rightarrow +\infty$.

En pratique, la Définition I.24 n'est pas très sympathique à manipuler si on souhaite calculer $\|A\|$. Heureusement, on dispose d'un résultat permettant de ramener le calcul de cette norme à un calcul de valeurs propres :

¹En VF https://fr.wikipedia.org/wiki/Norme_matricielle ou en VA (plus complète) https://en.wikipedia.org/wiki/Matrix_norm

Proposition I.28. Soit $A \in \mathcal{M}_{M,N}(\mathbb{R})$. Alors :

$$\|A\| = \sqrt{\rho(A^\top A)}.$$

I.I.4 Matrices symétriques et antisymétriques

Définition I.29. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice carrée. On dit que

- A est **SYMÉTRIQUE**, si $A^\top = A$.
- A est **ANTISYMÉTRIQUE**, si $A^\top = -A$.

Exercice I.30. Pour toute matrice $A \in \mathcal{M}_{M,N}(\mathbb{R})$, montrer que les matrices $A^\top A \in \mathcal{M}_N(\mathbb{R})$ et $AA^\top \in \mathcal{M}_M(\mathbb{R})$ sont symétriques.

Exercice I.31. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice antisymétrique. Montrer que, pour tout $x \in \mathbb{R}^N$, $\langle Ax, x \rangle = 0$.

Exercice I.32. Soit $A \in \mathcal{M}_N(\mathbb{R})$ quelconque. Montrer que $A + A^\top$ est symétrique, et que $A - A^\top$ est antisymétrique.

Proposition I.33. Toute matrice $A \in \mathcal{M}_N(\mathbb{R})$ peut se décomposer comme la somme d'une matrice symétrique et d'une matrice antisymétrique. En effet :

$$(\forall A \in \mathcal{M}_N(\mathbb{R})) \quad A = \underbrace{\frac{A + A^\top}{2}}_{\text{symétrique}} + \underbrace{\frac{A - A^\top}{2}}_{\text{antisymétrique}}.$$

Remarque I.34. On peut en fait même montrer que la matrice symétrique $\frac{A+A^\top}{2}$ est la projection orthogonale de A sur le sous-espace vectoriel des matrices symétriques.

Théorème I.35 (Théorème spectral). Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique. Alors il existe

- une matrice diagonale réelle $D \in \mathcal{M}_N(\mathbb{R})$
 - une matrice inversible $U \in \mathcal{M}_N(\mathbb{R})$ telle que $U^{-1} = U^\top$ (une matrice orthogonale, donc)
- telles que $A = U^\top DU$.

En particulier, toute matrice symétrique de $\mathcal{M}_N(\mathbb{R})$ est diagonalisable dans \mathbb{R} , et admet N valeurs propres réelles (en comptant les éventuelles multiplicités). Ce qu'il y a d'avantageux avec les matrices symétriques, c'est que de nombreuses propriétés/définitions/quantités associées aux matrices en général peuvent se réexprimer simplement en fonctions des valeurs propres. Et comme les valeurs propres sont calculables², c'est très utile en pratique. Par exemple :

²Facile à la main pour $N = 2$, faisable à la main pour $N = 3$, pour le reste on laisse un programme numérique s'en charger (approximativement).

Proposition I.36. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique. Alors sa norme d'opérateur est égale au rayon spectral :

$$\|A\| = \rho(A).$$

Remarque I.37. « La norme d'opérateur est égale au rayon spectral » est faux en général, puisque cela s'applique seulement aux matrices symétriques. Pour une matrice générale, c'est la Proposition I.28 qui s'applique. Pour s'en rendre compte, considérons par exemple

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \text{ telle que } A^\top A = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

On voit que $\text{spec}(A^\top A) = \{1, 0\}$, donc on déduit de la Proposition I.28 que $\|A\| = 1$. Pour autant, $\text{spec}(A) = \{0\}$ (immédiat puisque A est triangulaire avec des zéros sur la diagonale) donc $\rho(A) = 0$. Ici, la norme d'opérateur est bien différente du rayon spectral.

Puisque les matrices symétriques ont des valeurs propres réelles, on introduit deux notations qui nous seront utiles par la suite :

Définition I.38. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique. On note

- $\lambda_{\min}(A) \in \mathbb{R}$ la plus petite valeur propre de A ,
- $\lambda_{\max}(A) \in \mathbb{R}$ la plus grande valeur propre de A .

Proposition I.39. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique. Alors

$$(\forall x \in \mathbb{R}^N) \quad \lambda_{\min}(A)\|x\|^2 \leq \langle Ax, x \rangle \leq \lambda_{\max}(A)\|x\|^2.$$

Remarque I.40. Quelques commentaires sur l'utilité de ces deux inégalités :

- Si on veut borner supérieurement la quantité $\langle Ax, x \rangle$, on peut utiliser l'inégalité de Cauchy-Schwarz puis la définition de norme d'opérateur pour écrire :

$$\langle Ax, x \rangle \leq \|Ax\|\|x\| \leq \|A\|\|x\|^2.$$

Or on a toujours $\lambda_{\max}(A) \leq \|A\|$ donc le résultat de la proposition est plus précis en général.

- Cette borne inférieure est la « seule » inégalité classique dont on dispose pour borner inférieurement des quantités faisant intervenir une matrice.

Remarque I.41 (Inégalité de l'ellipse). Lorsque la matrice symétrique A est également à valeurs propres positives, on peut visualiser cette inégalité comme le fait de chercher les cercles inscrit et circonscrit à une ellipse. Considérons par exemple dans \mathbb{R}^2 la matrice $A = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$, telle que $\lambda_{\min}(A) = 1$ et $\lambda_{\max}(A) = 4$. Considérons les fonctions de $\mathbb{R}^2 \rightarrow \mathbb{R}$ suivantes

$$f_{\text{circ}} : (x_1, x_2) \mapsto x_1^2 + x_2^2, \quad f_{\text{ell}} : (x_1, x_2) \mapsto x_1^2 + 4x_2^2, \quad f_{\text{insc}} : (x_1, x_2) \mapsto 4x_1^2 + 4x_2^2.$$

La Proposition I.39 ne dit rien d'autre que le fait que $f_{circ}(x_1, x_2) \leq f_{ell}(x_1, x_2) \leq f_{insc}(x_1, x_2)$. L'ordre entre ces fonctions peut se voir clairement lorsque on trace leur graphe (cf Figure I.1).

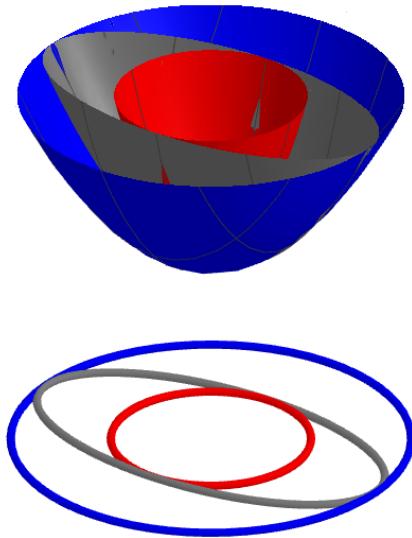


FIGURE I.1 – Inégalité de la Proposition I.39 pour $A = \begin{pmatrix} 1 & 0 \\ 0 & 4 \end{pmatrix}$.

Attention toutefois à bien garder en tête que la Proposition I.39 est encore vraie lorsque $\lambda_{\min}(A) < 0$! Dans ce cas, cette histoire d'ellipses ne tient plus puisque la fonction quadratique associée à A est dégénérée, et ses courbes de niveaux ne sont plus des ellipses mais des hyperboles (voir Figure I.2).

I.I.5 Matrices semi-définies positives et définies positives

I.I.5.i) La théorie

Définition I.42. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice. On dit que :

- A est **SEMI-DÉFINIE POSITIVE**, et on note $A \succeq 0$, si

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

- A est **DÉFINIE POSITIVE**, et on note $A \succ 0$, si

$$(\forall x \in \mathbb{R}^N \setminus \{0\}) \quad \langle Ax, x \rangle > 0.$$

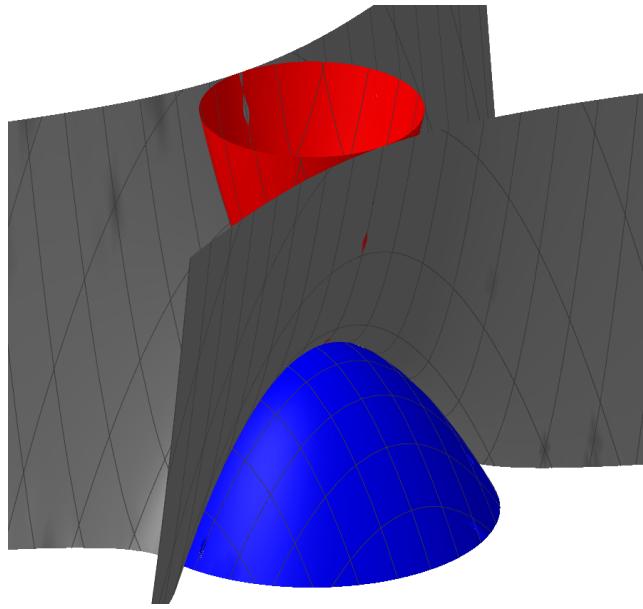


FIGURE I.2 – Inégalité de la Proposition I.39 pour $A = \begin{pmatrix} -1 & 0 \\ 0 & 4 \end{pmatrix}$.

Remarque I.43 (Matrice semi-définie positive vs. coefficients positifs). La notion de matrice semi-définie positive est parfois confondue avec la notion de « matrice-dont-les-coefficients-sont-positifs », or ces deux notions n'ont rien en commun. Par exemple, la matrice

$$\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad (\text{I.1})$$

possède un coefficient négatif, néanmoins elle est bien semi-définie positive puisque

$$(\forall (x, y) \in \mathbb{R}^2) \quad \langle \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}, \begin{pmatrix} x \\ y \end{pmatrix} \rangle = \langle \begin{pmatrix} -y \\ x \end{pmatrix}, \begin{pmatrix} x \\ y \end{pmatrix} \rangle = -yx + xy = 0 \geq 0.$$

D'un autre côté, la matrice

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

est clairement à coefficient positifs, néanmoins on peut montrer qu'elle n'est pas une matrice semi-définie positive. (On reverra cet exemple plus tard)

Remarque I.44 (Matrice semi-définie positive et valeurs propres). Une autre confusion fréquente est la suivante :

« Une matrice est semi-définie positive si et seulement si ses valeurs propres sont positives »,

voire également :

« Une matrice est définie positive si et seulement si ses valeurs propres sont strictement positives ».

Ces deux énoncés sont **faux** en général. Rappelons par exemple qu'une matrice carrée n'admet pas nécessairement de valeurs propres, c'est le cas de la matrice (I.1) qui n'admet aucune valeur propre réelle, mais qui pourtant est bien semi-définie positive. Par contre que ces énoncés sont vrais **si la matrice en question est symétrique** :

Proposition I.45. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique. Alors on a les équivalences suivantes :

- i) A est semi-définie positive $\Leftrightarrow \lambda_{\min}(A) \geq 0$.
- ii) A est définie positive $\Leftrightarrow \lambda_{\min}(A) > 0 \Leftrightarrow A$ est semi-définie positive et inversible.

Exercice I.46. Soit $A \in \mathcal{M}_{M,N}(\mathbb{R})$, montrer que :

- 1) les matrices $A^\top A \in \mathcal{M}_N(\mathbb{R})$ et $AA^\top \in \mathcal{M}_M(\mathbb{R})$ sont symétriques semi-définies positives ;
- 2) $A^\top A$ est définie positive si et seulement si A est injective ;
- 3) AA^\top est définie positive si et seulement si A est surjective.

Et pour les matrices non symétriques ? Eh bien nous pouvons toujours nous *ramener* aux matrices symétriques, grâce au résultat suivant :

Proposition I.47. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice carrée. Alors :

A est (semi-) définie positive si et seulement si $\frac{A^\top + A}{2}$ est (semi-) définie positive .

En pratique, pour une matrice carrée A quelconque, il suffit donc de vérifier le signe des valeurs propres de la matrice symétrique $\frac{A^\top + A}{2}$.

Exemple I.48. Si on considère la matrice triangulaire $A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$, on voit que $\text{spec}(A) = \{0\}$. Mais on ne peut pas en déduire immédiatement que A est semi-définie positive, puisque elle n'est pas symétrique ! Par contre on peut calculer $\frac{A^\top + A}{2} = \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix}$, dont l'ensemble des valeurs propres est $\{\pm 1/2\}$. Puisque l'une des valeurs propres est négative, on en déduit que A n'est pas une matrice semi-définie positive.

Donc, moralement, la question de la positivité (resp. définie positivité) d'une matrice peut toujours se ramener à celle de la positivité (resp. stricte positivité) de toutes les valeurs propres d'une matrice. Mais que se passe-t-il lorsque certaines de ces valeurs propres

sont négatives ? Si elles le sont toutes, on parle de matrice semi-définie négative, sinon on parle de matrice indéfinie :

Définition I.49. On dit que $A \in \mathcal{M}_N(\mathbb{A})$ est :

- **SEMI-DÉFINIE NÉGATIVE**, et on note $A \preceq 0$, si $-A$ est semi-définie positive :

$$(\forall x \in \mathbb{R}^N) \quad \langle Ax, x \rangle \leq 0.$$

- **DÉFINIE NÉGATIVE**, et on note $A < 0$, si $-A$ est définie positive :

$$(\forall x \in \mathbb{R}^N \setminus \{0\}) \quad \langle Ax, x \rangle < 0.$$

- **INDÉFINIE** si elle n'est ni semi-définie positive ni semi-définie négative. Autrement dit, si

$$(\exists x_1, x_2 \in \mathbb{R}^N) \quad \langle Ax_1, x_1 \rangle < 0 \quad \text{et} \quad \langle Ax_2, x_2 \rangle > 0.$$

Exemple I.50. Il peut être intéressant de visualiser ces propriétés d'une matrice A en regardant le graphe de la fonction quadratique associée $q_A : x \mapsto \langle Ax, x \rangle$. Comme on peut le voir dans la figure I.3, les formes quadratiques définies positives montent à l'infini dans toutes les directions. Lorsque A est semi-définie positive mais pas définie positive, cela veut dire qu'il y a un noyau non nul, ce qui se traduit par des directions où la forme quadratique est constante. Lorsque A est non définie, la forme quadratique peut tendre vers $+\infty$ ou $-\infty$, selon la direction dans laquelle on va. Dans ce cas on parle souvent de *point selle*, qui est une notion que l'on reverra bientôt.

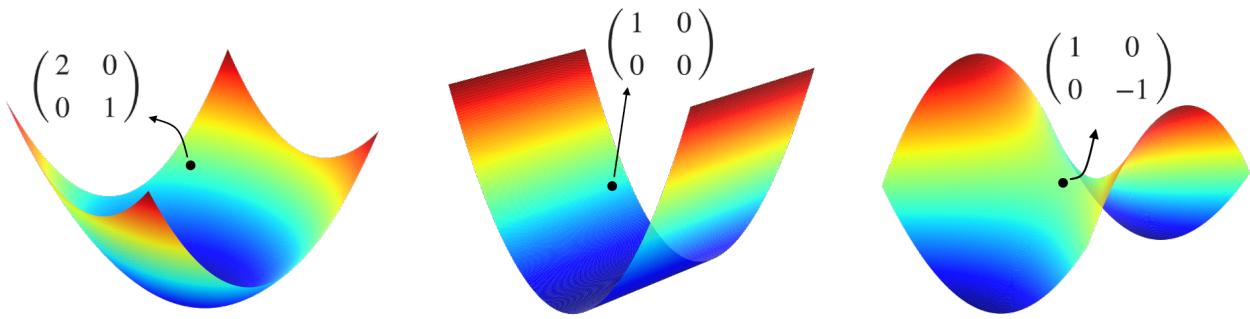


FIGURE I.3 – Formes quadratiques respectivement associées à une matrice définie positive, semi-définie positive et non définie.

I.I.5.ii) La pratique

Un réflexe naturel pour déterminer la positivité d'une matrice symétrique est de calculer ses valeurs propres, puis de simplement vérifier leur signe. Or, calculer les valeurs

propres, ce n'est pas facile lorsque la dimension dépasse 3 (et déjà pour $N = 3$ ce n'est pas très sympathique).

Mais en réalité nous n'avons pas besoin de calculer les valeurs propres ; tout ce dont on a besoin est leur *signe*. Par exemple, pour les matrices 2×2 :

Exercice I.51. (Positivité d'une matrice symétrique 2×2) Soit $A \in \mathcal{M}_2(\mathbb{R})$ une matrice symétrique. Montrer que A est semi-définie positive (resp. définie positive) si et seulement si sa trace et son déterminant sont positifs (resp. strictement positifs).

Exercice I.52 (Matrices semi-définies positives et définies positives). Déterminer la nature des matrices suivantes (définie positive, semi-définie positive ou non définie) :

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 1 \\ 1 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 4 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 2 & 0 \\ -2 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & 3 \\ -1 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & -2 \\ 0 & 1 \end{pmatrix}$$

Ce critère ne vaut évidemment que pour les matrices de taille 2. Pour des matrices plus grandes, on dispose en fait d'un critère plus général, qui passe par le calcul de déterminants de certaines sous-matrices :

Théorème I.53 (Critère de Sylvester). Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique. Alors :

- i) A est semi-définie positive si et seulement si tous ses mineurs principaux sont positifs.
- ii) A est définie positive si et seulement si tous ses mineurs principaux sont strictement positifs.

La sous-section suivante présente la notion de mineurs principaux, si vous ne l'avez jamais vue.

I.I.5.iii) Mineurs principaux

Définition I.54. Soit $A \in \mathcal{M}_N(\mathbb{R})$ et $I \subsetneq \{1, \dots, N\}$. On note A_I la sous-matrice de A obtenue en lui retirant ses i -ème ligne et i -ème colonne, pour tout $i \in I$. On dit que $A_I \in \mathcal{M}_{N-|I|}(\mathbb{R})$ est une sous-matrice **PRINCIPALE**.

Exercice I.55. Listons toutes les sous-matrices principales de la matrice

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}.$$

Pour commencer, il n'y a qu'une sous-matrice principale de taille 3, qui est A elle-même. On l'obtient avec A_I en prenant $I = \emptyset$. Ensuite viennent les sous-matrices de taille 2, qui

s'obtiennent en retirant i -ème ligne et i -ème colonne pour $i = 1, 2, 3$:

$$A_I = \underbrace{\begin{pmatrix} 5 & 6 \\ 8 & 9 \end{pmatrix}}_{I=\{1\}}, \quad \underbrace{\begin{pmatrix} 1 & 3 \\ 7 & 9 \end{pmatrix}}_{I=\{2\}}, \quad \underbrace{\begin{pmatrix} 1 & 2 \\ 4 & 5 \end{pmatrix}}_{I=\{3\}}.$$

Enfin, les sous-matrices de taille 1, qui s'obtiennent en retirant deux lignes et deux colonnes, et qui correspondent aux éléments diagonaux :

$$A_I = \underbrace{\begin{pmatrix} 9 \end{pmatrix}}_{I=\{1,2\}}, \quad \underbrace{\begin{pmatrix} 5 \end{pmatrix}}_{I=\{1,3\}}, \quad \underbrace{\begin{pmatrix} 1 \end{pmatrix}}_{I=\{2,3\}}.$$

Exercice I.56. Listons toutes les sous-matrices principales de la matrice

$$A = \begin{pmatrix} 01 & 02 & 03 & 04 \\ 05 & 06 & 07 & 08 \\ 09 & 10 & 11 & 12 \\ 13 & 14 & 15 & 16 \end{pmatrix}.$$

Pour commencer, il n'y a qu'une sous-matrice principale de taille 4, qui est A elle-même. Ensuite les sous-matrices de taille 3, qui s'obtiennent en retirant i -ème ligne et i -ème colonne pour $i = 1..4$:

$$A_I = \underbrace{\begin{pmatrix} 06 & 07 & 08 \\ 10 & 11 & 12 \\ 14 & 15 & 16 \end{pmatrix}}_{I=\{1\}}, \quad \underbrace{\begin{pmatrix} 01 & 03 & 04 \\ 09 & 11 & 12 \\ 13 & 15 & 16 \end{pmatrix}}_{I=\{2\}}, \quad \underbrace{\begin{pmatrix} 01 & 02 & 04 \\ 05 & 06 & 08 \\ 13 & 14 & 16 \end{pmatrix}}_{I=\{3\}}, \quad \underbrace{\begin{pmatrix} 01 & 02 & 03 \\ 05 & 06 & 07 \\ 09 & 10 & 11 \end{pmatrix}}_{I=\{4\}}.$$

Ensuite les sous-matrices de taille 2, qui s'obtiennent en retirant une paire de lignes/colonnes à A . On peut également les obtenir en retirant UNE ligne/colonne aux sous-matrices principales de taille 3 :

$$A_I = \underbrace{\begin{pmatrix} 11 & 12 \\ 15 & 16 \end{pmatrix}}_{I=\{1,2\}}, \quad \underbrace{\begin{pmatrix} 06 & 08 \\ 14 & 16 \end{pmatrix}}_{I=\{1,3\}}, \quad \underbrace{\begin{pmatrix} 06 & 07 \\ 10 & 11 \end{pmatrix}}_{I=\{1,4\}}, \quad \underbrace{\begin{pmatrix} 01 & 04 \\ 13 & 16 \end{pmatrix}}_{I=\{2,3\}}, \quad \underbrace{\begin{pmatrix} 01 & 03 \\ 09 & 11 \end{pmatrix}}_{I=\{2,4\}}, \quad \underbrace{\begin{pmatrix} 01 & 02 \\ 05 & 06 \end{pmatrix}}_{I=\{3,4\}}.$$

Enfin, les sous-matrices de taille 1, qui correspondent aux éléments diagonaux :

$$A_I = \underbrace{\begin{pmatrix} 16 \end{pmatrix}}_{I=\{1,2,3\}}, \quad \underbrace{\begin{pmatrix} 11 \end{pmatrix}}_{I=\{1,2,4\}}, \quad \underbrace{\begin{pmatrix} 06 \end{pmatrix}}_{I=\{1,3,4\}}, \quad \underbrace{\begin{pmatrix} 01 \end{pmatrix}}_{I=\{2,3,4\}}.$$

L'ensemble des mineurs principaux d'une matrice correspond simplement à l'ensemble des déterminants de toutes ses sous-matrices principales :

Définition I.57. Soit $A \in \mathcal{M}_N(\mathbb{R})$. On définit l'ensemble de ses **MINEURS PRINCIPAUX** par

$$\{\det(A_I) : I \subsetneq \{1, \dots, N\}\} \subset \mathbb{R}.$$

I.II Rappels et compléments de calcul différentiel

La notation $o(g(h))$ désigne une fonction signifie qu'il existe une fonction $\varepsilon : \mathbb{R}^N \rightarrow \mathbb{R}$ telle que $\lim_{h \rightarrow 0} \varepsilon(h) = 0$, et qui permette d'écrire le reste sous la forme $o(g(h)) = g(h)\varepsilon(h)$.

Étant donné une fonction $F : U \subset \mathbb{R}^N \rightarrow \mathbb{R}^M$, on notera $F_1, \dots, F_M : U \rightarrow \mathbb{R}$ les fonctions qui vérifient

$$(\forall x \in U) \quad F(x) = (F_1(x), \dots, F_M(x)).$$

Une autre façon d'écrire ceci est de poser $F_i(x) = \langle F(x), e_i \rangle$ où e_i est le i -ème vecteur de la base canonique de \mathbb{R}^M .

I.II.1 Différentielle

Définition I.58 (Différentielle). Soit $U \subset \mathbb{R}^N$ un ouvert et $F : U \rightarrow \mathbb{R}^M$ une application. Soit $x \in U$. On dit que F est **DIFFÉRENTIABLE** au point x s'il existe une application linéaire $u \in \mathcal{L}(\mathbb{R}^N; \mathbb{R}^M)$ telle que pour tout $h \in \mathbb{R}^N$ t.q. $x + h \in U$,

$$F(x + h) = F(x) + u(h) + o(\|h\|).$$

Lorsque u existe, elle est unique ; on la note $u = DF(x)$.

Si l'application $x \mapsto DF(x)$ est définie sur tout U , et y est continue, on dit alors que F est de classe C^1 sur U et on note $F \in C^1(U)$.

Définition I.59 (Dérivée directionnelle). Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ et $d \in \mathbb{R}^N \setminus \{0\}$. On dit que f admet une **DÉRIVÉE DIRECTIONNELLE** dans la direction d , au point x , si l'application $t \in \mathbb{R} \mapsto f(x + td)$ est dérivable en 0. Si c'est le cas, on note cette dérivée

$$\frac{\partial f}{\partial d}(x) := \lim_{t \rightarrow 0} \frac{f(x + td) - f(x)}{t}.$$

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

$$\frac{\partial f}{\partial x_i}(x) := \lim_{t \rightarrow 0} \frac{f(x + te_i) - f(x)}{t} = \lim_{t \rightarrow 0} \frac{f(x_1, \dots, x_{i-1}, x_i + t, x_{i+1}, \dots, x_N) - f(x_1, \dots, x_N)}{t}.$$

Remarque I.60 (Matrice Jacobienne). Toute application linéaire $u \in \mathcal{L}(\mathbb{R}^N, \mathbb{R}^M)$ peut être représentée par une matrice $A \in \mathcal{M}_{M,N}(\mathbb{R})$ telle que $u(x)$ soit égale au produit matriciel Ax . Plus précisément, cette matrice A est la matrice représentant u dans la base canonique. Dans le cas de la différentielle $DF(x)$, sa matrice associée est la matrice JACOBIENNE, que l'on note $JF(x)$. Au vu de la définition précédente, cette matrice vérifie

$$F(x + h) = F(x) + JF(x)h + o(\|h\|).$$

On se rappelle en général de la matrice Jacobienne comme étant « la matrice des dérivées partielles » de F . C'est effectivement le cas, comme le prouve la prochaine Proposition :

Proposition I.61 (Jacobienne et dérivées partielles). Soit $U \subset \mathbb{R}^N$ un ouvert et $F : U \rightarrow \mathbb{R}^M$ une fonction différentiable en $x \in U$. Alors :

- i) Pour tout $i = 1, \dots, M$, F_i admet des dérivées directionnelles en toute direction au point x . En particulier, elle admet des dérivées partielles en x .
- ii) Les coefficients de la matrice Jacobienne $JF(x)$ sont des dérivées partielles en x :

$$JF(x) = \begin{bmatrix} \frac{\partial F_1}{\partial x_1}(x) & \cdots & \frac{\partial F_1}{\partial x_N}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial F_M}{\partial x_1}(x) & \cdots & \frac{\partial F_M}{\partial x_N}(x) \end{bmatrix} = \begin{bmatrix} JF_1(x) \\ \vdots \\ JF_M(x) \end{bmatrix}$$

Remarque I.62 (Vecteur Gradient). Si $f : \mathbb{R}^N \rightarrow \mathbb{R}^1$ (on insiste sur le fait que $M = 1$) est différentiable en x , alors $Jf(x) \in \mathcal{M}_{1,N}(\mathbb{R})$ est un vecteur ligne (et $Df(x)$ est une forme linéaire). Sa transposée est donc identifiable à un vecteur (colonne), que l'on appelle le GRADIENT de f en x : $\nabla f(x) = Jf(x)^T$.

Proposition I.63. Si $f : U \subset \mathbb{R}^N \rightarrow \mathbb{R}$ est différentiable en $x \in U$, alors :

- i) Elle admet des dérivées directionnelles en toute direction au point x (et en particulier, des dérivées partielles).
- ii) Le gradient de f en x s'écrit

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x) \\ \vdots \\ \frac{\partial f}{\partial x_N}(x) \end{bmatrix}$$

- iii) On a la relation suivante entre différentielle, gradient, dérivée directionnelle et dérivée partielle :

$$(\forall d \in \mathbb{R}^N) \quad Df(x)(d) = \frac{\partial f}{\partial d}(x) = \langle \nabla f(x), d \rangle = \sum_{i=1}^N \frac{\partial f}{\partial x_i}(x) d_i.$$

Remarque I.64 (Calcul du gradient). Pour calculer le gradient de f au point x , il n'est pas toujours nécessaire de calculer explicitement toutes les dérivées partielles. Une autre méthode consiste à établir un développement limité de f sous la forme suivante :

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

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

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

Théorème I.65 (Règles de calcul).

i) Soient $F, G : U \subset \mathbb{R}^N \rightarrow \mathbb{R}^M$ deux fonctions différentiables en $x \in U$. Alors $F + G$ est différentiable en x , et

$$D(F + G)(x) = DF(x) + DG(x) \text{ et } J(F + G)(x) = JF(x) + JG(x).$$

ii) Soient $G : \mathbb{R}^N \rightarrow \mathbb{R}^M$ et $F : \mathbb{R}^M \rightarrow \mathbb{R}^P$, telles que G soit différentiable en x et F soit différentiable en $G(x)$. Alors $F \circ G : \mathbb{R}^N \rightarrow \mathbb{R}^P$ est différentiable en x , et

$$D(F \circ G)(x) = DF(G(x)) \circ DG(x) \quad \text{et} \quad \underbrace{J(F \circ G)(x)}_{\in \mathcal{M}_{P,N}(\mathbb{R})} = \underbrace{JF(G(x))}_{\in \mathcal{M}_{P,M}(\mathbb{R})} \underbrace{JG(x)}_{\in \mathcal{M}_{M,N}(\mathbb{R})}.$$

iii) Soient $G : \mathbb{R}^N \rightarrow \mathbb{R}^M$ et $f : \mathbb{R}^M \rightarrow \mathbb{R}$, telles que G soit différentiable en x et f soit différentiable en $G(x)$. Alors $f \circ G : \mathbb{R}^N \rightarrow \mathbb{R}$ est différentiable en x , et

$$\underbrace{\nabla(f \circ G)(x)}_{\in \mathbb{R}^N} = \underbrace{JG(x)^\top}_{\in \mathcal{M}_{N,M}(\mathbb{R})} \underbrace{\nabla f(G(x))}_{\in \mathbb{R}^M}.$$

Exemple I.66. Soit $f : \mathbb{R} \rightarrow \mathbb{R}$. Alors $\nabla f(x) = f'(x)$.

Exemple I.67. Soit $f(x) = \frac{1}{2}\|x\|^2$, alors $\nabla f(x) = x$ et $Df(x) = x^T$.

Exemple I.68. Si $F : \mathbb{R}^N \rightarrow \mathbb{R}^M$ est constante, alors $DF(x) = 0$.

Exemple I.69. Si $F : \mathbb{R}^N \rightarrow \mathbb{R}^M$ est linéaire, alors $DF(x) = F$.

Exercice I.70 (Dériver la trace). Soit $f : \mathcal{M}_N(\mathbb{R}) \rightarrow \mathbb{R}$ définie par $f(X) = \text{tr}(X)$.

1) Calculer $Df(X)$, pour $X \in \mathcal{M}_N(\mathbb{R})$.

2) On munit $\mathcal{M}_N(\mathbb{R})$ du produit scalaire suivant (on admet que c'est un produit scalaire) :

$$(\forall X, Y \in \mathcal{M}_N(\mathbb{R})) \quad \langle \langle X, Y \rangle \rangle = \text{tr}(X^\top Y).$$

Calculer $\nabla f(X)$.

Exercice I.71 (Gradient d'une composée). Soit $g : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiable, et $f(x) = g(x)_+^2$, où la notation x_+ veut dire $\max\{0, x\}$ (on parle de *partie positive*). Calculer $\nabla f(x)$. même question avec $f(x) = g(x)^2$.

Exemple I.72. Soit $f(x) = g(Ax + b)$ où $A \in \mathcal{M}_{M,N}(\mathbb{R})$ et $g : \mathbb{R}^M \rightarrow \mathbb{R}$ est différentiable. Alors $\nabla f(x) = A^T \nabla g(Ax + b)$.

Exemple I.73. Si $f(x) = \frac{1}{2}\|Ax - b\|^2$, alors $\nabla f(x) = A^T(Ax - b)$.

On termine avec un résultat qui n'est pas central dans ce cours, mais que l'on utilisera par la suite dans les preuves :

Proposition I.74 (Théorème de Taylor-Lagrange, ordre 1). Soit $a \in \mathbb{R}^N$, $U = \mathbb{B}(a, R)$ une boule ouverte de \mathbb{R}^N , et $f : U \rightarrow \mathbb{R}$ de classe $C^1(U)$. Alors, pour tout $x \in U$, il existe $z \in]a, x[$ tel que

$$f(x) = f(a) + \langle \nabla f(z), x - a \rangle.$$

I.II.2 Différentielle seconde

Définition I.75 (Différentielle seconde). Soit U un ouvert de \mathbb{R}^N et $F : U \rightarrow \mathbb{R}^M$. On dit que F est deux fois différentiable en $x \in U$ si F est différentiable sur U , et s'il existe une application bilinéaire symétrique $b \in \mathcal{B}(\mathbb{R}^N, \mathbb{R}^N; \mathbb{R}^M)$ telle que

$$(\forall h \in U - x) \quad F(x + h) = F(x) + DF(x)(h) + \frac{1}{2}b(h, h) + o(\|h\|^2).$$

Dans ce cas b est uniquement définie, et c'est la différentielle seconde de F en x , notée $D^2F(x)$. Si l'application $x \mapsto D^2F(x)$ existe et est continue sur U , on note $F \in C^2(U)$.

Proposition I.76 (La différentielle de la différentielle). Soit $F : U \subset \mathbb{R}^N \rightarrow \mathbb{R}^M$ deux fois différentiable en $x \in U$. Alors

$$(\forall h, k \in \mathbb{R}^N) \quad D^2F(x)(h, k) = D(DF)(x)(h)(k).$$

Remarque I.77 (Matrice hessienne). Pour toute application bilinéaire $b \in \mathcal{B}(\mathbb{R}^N, \mathbb{R}^N; \mathbb{R})$ il existe une unique matrice $B \in \mathbb{R}^{N \times N}$ telle que $b(x, y) = \langle Bx, y \rangle$. Cela revient à dire que $B_{ij} = b(e_i, e_j)$. Dans le cas de la différentielle seconde $D^2f(x)$ d'une fonction f de $\mathbb{R}^N \rightarrow \mathbb{R}$, la matrice associée est la matrice HESSIENNE, notée $\nabla^2f(x)$, et qui vérifie les propriétés suivantes :

Proposition I.78. Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ une fonction deux fois différentiable en $x \in U$. Alors :

- i) (Symétrie) $\nabla^2f(x)$ est une matrice symétrique.
- ii) (Matrice des dérivées partielles seconde) $\nabla^2f(x) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right)_{ij}$.
- iii) (Jacobienne du gradient) $\nabla^2f(x) = J(\nabla f)(x)$.
- iv) (Taylor ordre 2) $(\forall h \in \mathbb{R}^N) \quad f(x + h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2}\langle \nabla^2f(x)h, h \rangle + o(\|h\|^2)$.

Théorème I.79 (Règles de calcul).

- i) Soient $F, G : U \subset \mathbb{R}^N \rightarrow \mathbb{R}^M$ deux fonctions deux fois différentiables en $x \in U$. Alors $F + G$ est deux fois différentiable en x , et

$$D^2(F + G)(x) = D^2F(x) + D^2G(x).$$

ii) Soient $G : \mathbb{R}^N \rightarrow \mathbb{R}^M$ et $F : \mathbb{R}^M \rightarrow \mathbb{R}^P$, telles que G soit deux fois différentiable en x et F soit deux fois différentiable en $G(x)$. Alors $F \circ G : \mathbb{R}^N \rightarrow \mathbb{R}^P$ est deux fois différentiable en x , et

$$D^2(F \circ G)(x) = D^2F(G(x)) \circ (DG(x) \otimes DG(x)) + DF(G(x)) \circ D^2G(x),$$

autrement dit, pour tout $h, k \in \mathbb{R}^N$:

$$D^2(F \circ G)(x)(h, k) = D^2F(G(x))(DG(x)(h), DG(x)(k)) + DF(G(x))(D^2G(x)(h, k)).$$

iii) Soient $G : \mathbb{R}^N \rightarrow \mathbb{R}^M$ et $f : \mathbb{R}^M \rightarrow \mathbb{R}$, telles que G soit deux fois différentiable en x et f soit deux fois différentiable en $G(x)$. Alors $f \circ G : \mathbb{R}^N \rightarrow \mathbb{R}$ est deux fois différentiable en x , et

$$\nabla^2(f \circ G)(x) = JG(x)^\top \nabla^2 f(G(x)) JG(x) + \sum_{i=1}^M \frac{\partial f}{\partial x_i}(G(x)) \nabla^2 G_i(x).$$

Remarque I.80. La règle de calcul un peu barbare pour la différentielle seconde de la composition est à rapprocher de celle que l'on retrouve facilement pour la dérivée seconde de la composée de deux fonctions réelles :

$$(f \circ g)''(x) = (f' \circ g \cdot g')'(x) = f''(g(x))g'(x)g'(x) + f'(g(x))g''(x).$$

Exemple I.81. Soit $f : \mathbb{R} \rightarrow \mathbb{R}$. Alors $\nabla^2 f(x) = f''(x)$.

Exemple I.82. Soit $f(x) = \frac{1}{2}\|x\|^2$, alors $\nabla^2 f(x) = Id$ et $D^2 f(x) = \langle \cdot, \cdot \rangle$.

Exemple I.83. Si $F : \mathbb{R}^N \rightarrow \mathbb{R}^M$ est linéaire ou affine alors $D^2 F(x) = 0$.

Exemple I.84. Soit $f(x) = g(Ax + b)$ où $A \in \mathcal{M}_{M,N}(\mathbb{R})$ et $g : \mathbb{R}^M \rightarrow \mathbb{R}$ est deux fois différentiable. Alors $\nabla^2 f(x) = A^T \nabla^2 g(Ax + b)A$.

Exemple I.85. Si $f(x) = \frac{1}{2}\|Ax - b\|^2$, alors $\nabla^2 f(x) = A^T A$.

Exemple I.86. Si $f(x) = \langle Ax, x \rangle$, alors $\nabla^2 f(x) = (A + A^T)/2$. Si A est symétrique on obtient $\nabla^2 f(x) = A$.

Proposition I.87 (Théorème de Taylor-Lagrange, ordre 2). Soit $a \in \mathbb{R}^N$, $U = \mathbb{B}(a, R)$ une boule ouverte de \mathbb{R}^N , et $f : U \rightarrow \mathbb{R}$ de classe $C^2(U)$. Alors, pour tout $x \in U$, il existe $z \in]a, x[$ tel que

$$f(x) = f(a) + \langle \nabla f(a), x - a \rangle + \frac{1}{2} \langle \nabla^2 f(z)(x - a), x - a \rangle.$$

I.II.3 Fonctions quadratiques

Définition I.88. (Fonction quadratique) On dira qu'une fonction $f : \mathbb{R}^N \rightarrow \mathbb{R}$ est **QUADRATIQUE** si elle peut s'écrire sous la forme

$$f(x) = \langle Ax, x \rangle + \langle b, x \rangle + c,$$

où $A \in \mathcal{M}_N(\mathbb{R})$, $b \in \mathbb{R}^N$ et $c \in \mathbb{R}$.

Remarque I.89. Les fonctions quadratiques sont des polynômes de degré 2 en les variables x_1, \dots, x_N . En effet, en notant a_{ij} et b_i les coefficients de A et b , on peut écrire

$$f(x) = \sum_{i=1}^N \sum_{j=1}^N a_{ij} x_j x_i + \sum_{i=1}^N b_i x_i + c.$$

Exemple I.90. Les fonctions quadratiques de \mathbb{R} dans \mathbb{R} sont exactement les « fonctions du second degré » abondamment étudiées au lycée : $f(x) = ax^2 + bx + c$.

Exemple I.91. $f(x, y) = 2x^2 + y^2 - xy + 3x - 2$ est une fonction quadratique sur \mathbb{R}^2 .

Exemple I.92. $f(x, y) = 2x^2 + y^2 - xy^2 + 3x - 2$ n'est pas une fonction quadratique sur \mathbb{R}^2 car c'est un polynôme de degré 3.

Proposition I.93. Soit $f(x) = \langle Ax, x \rangle + \langle b, x \rangle + c$ une fonction quadratique sur \mathbb{R}^N . Alors

$$\nabla f(x) = (A + A^\top)x + b \quad \text{et} \quad \nabla^2 f(x) = A + A^\top.$$

En particulier, si A est symétrique, on a $\nabla f(x) = 2Ax + b$ et $\nabla^2 f(x) = 2A$.

Exercice I.94 (Moindre carrés). Soit $A \in \mathcal{M}_{M,N}(\mathbb{R})$, $y \in \mathbb{R}^M$, et $f : \mathbb{R}^N \rightarrow \mathbb{R}$ définie par

$$f(x) = \|Ax - y\|^2.$$

Montrer que f est une fonction quadratique, et calculer son gradient et sa Hessienne.

Chapitre II

Existence de minimiseurs et conditions d'optimalité



FIGURE II.1 – « *La nature agit toujours par les voies les plus courtes* », Pierre de Fermat (1657). « *Lorsqu'il arrive quelque changement dans la Nature, la quantité d'action, nécessaire pour ce changement, est la plus petite qu'il soit possible* », Pierre de Maupertuis (1756).

Dans ce chapitre, on considérera que $U \subset \mathbb{R}^N$ est un ouvert, et que $f : U \rightarrow \mathbb{R}$, et que $C \subset U$ est non vide. On s'intéresse au problème d'optimisation suivant :

$$(P_C) \quad \inf_{x \in C} f(x)$$

Ici, on dit que C est la CONTRAINTE de notre problème d'optimisation. Lorsque $C = \mathbb{R}^N$, on parle en général d'optimisation sans contrainte.

II.I Conditions d'optimalité et Principe de Fermat

On commence ce chapitre en définissant :

II.I.1 Un peu de vocabulaire

Définition II.1. Soit $C \subset \mathbb{R}^N$, et $f : C \rightarrow \mathbb{R}$.

- L'**INFIMUM** de f , noté $\inf_C f$, est défini par $\inf_C f := \inf\{f(x) \mid x \in C\} \in [-\infty, +\infty[$.
- Lorsque $\inf_C f \neq -\infty$, on dit que f est **MINORÉE** sur C .
- On dit que $\bar{x} \in C$ est un **MINIMISEUR** de f sur C , si $f(\bar{x}) = \inf_C f$. Autrement dit, si

$$(\forall x \in C) \quad f(\bar{x}) \leq f(x).$$

- On note $\operatorname{argmin}_C f \subset C$ l'ensemble des minimiseurs de f sur C :

$$\operatorname{argmin}_C f = \{\bar{x} \in C \mid f(\bar{x}) = \inf_C f\}.$$

Lorsqu'on sait qu'il existe un minimiseur, on dit que l'infimum est atteint, et au lieu d'*infimum* on parle en général plutôt de **MINIMUM**, que l'on note $\min_C f$.

Enfin, lorsque $C = \mathbb{R}^N$, on omet de le mentionner, et on parlera simplement d'infimum ($\inf f$), minimum ($\min f$), minimiseur ($\operatorname{argmin} f$).

Remarque II.2 (Vocabulaire et subtilités).

- Il arrive parfois que l'on parle de minimum, ou de $\min_C f$, sans savoir s'il existe un minimiseur. C'est un léger abus, qu'on essaiera d'éviter dans ce cours, mais que vous allez très certainement rencontrer ailleurs.
- Il y a une ambiguïté beaucoup plus problématique concernant le terme *minimum*, dont le sens est souvent confondu avec celui de *minimiseur*. Martelons donc ici que :
 - le minimum désigne la plus petite valeur que peut prendre une fonction,
 - minimiseur désigne un point en lequel la fonction atteint son minimum.

Encore une fois, on essaiera dans ce cours de bien faire la différence entre les deux, et il est probable que vous trouviez une utilisation différente de ces termes dans des livres.

- Au lieu de minimiseur, on emploiera parfois le terme de **GLOBAL**, par opposition avec la Définition II.5 à venir. Les deux termes sont légitimes, on utilisera l'un ou l'autre en fonction du contexte.

Exemple II.3. Voici quelques exemples typiques, que je vous conseille de toujours garder en tête lorsque vous posez des questions sur les minimiseurs/minimum d'une fonction. Faites un dessin pour vous convaincre !

- (Pas minorée) $f(x) = x$, ou $f(x) = \ln x$ ne sont pas minorées : $\inf f = -\infty$ et $\operatorname{argmin} f = \emptyset$.

- (Minorée, pas de minimiseur) $f(x) = e^x$ pour laquelle $\inf f = 0$ mais $\operatorname{argmin} f = \emptyset$. Même chose pour $f(x) = 1/x$ sur $]0, +\infty[$.
- (Minimiseur unique) $f(x) = x^2$ pour laquelle $\min f = 0$ et $\operatorname{argmin} f = \{0\}$.
- (Minimiseurs multiples mais en nombre fini) $f(x) = ((x - 1)(x + 1))^2$ pour laquelle $\min f = 0$ et $\operatorname{argmin} f = \{-1, +1\}$.
- (Ensemble infini de minimiseurs, mais discret) $f(x) = \cos(x)$ pour laquelle $\min f = -1$ et $\operatorname{argmin} f = -\pi + 2\pi\mathbb{Z}$.
- (Continuum de minimiseurs) $f(x, y) = x^2$ pour laquelle $\inf f = 0$ et $\operatorname{argmin} f = \{0\} \times \mathbb{R}$.

Exercice II.4 (Existence de minimiseurs). Les fonctions suivantes atteignent-elles leur minimum ?

- 1) $f(x) = \exp(-x)$ sur $C = \mathbb{R}_+$, puis $C = \mathbb{R}_-$.
- 2) $f(x) = \cos(\exp(x^2))$ sur $C = [0, 1]$.
- 3) $f(x) = -\|x\|^2$ sur la boule fermée $C = \mathbb{B}(0, 1)$.
- 4) $f(x, y) = x^6 \cos y + 2y^2$ sur $C = \mathbb{R}^2$.

Les notions introduites dans la Définition II.1 peuvent être déclinées localement :

Définition II.5 (Minimiseur local). Soit $f : C \subset \mathbb{R}^N \rightarrow \mathbb{R}$. On dit que $\bar{x} \in C$ est un **MINIMISEUR LOCAL** de f sur C si

$$(\exists R > 0)(\forall x \in \mathbb{B}(\bar{x}, R) \cap C) \quad f(\bar{x}) \leq f(x).$$

Lorsque $C = \mathbb{R}^N$, on omettra de le mentionner, et on dira simplement que \bar{x} est un minimiseur local de f .

Remarque II.6. On peut reformuler la Définition II.5 ainsi : \bar{x} est un minimiseur local de f sur C si il existe un voisinage U de \bar{x} tel que \bar{x} soit un minimiseur (global) de f sur $C \cap U$.

Exemple II.7. L'existence d'un minimiseur local ne prédétermine en rien l'existence de minimiseurs globaux. Pire, on peut même avoir une fonction non minorée, comme par exemple la fonction polynomiale $f(x) = x(x - 1)(x + 1)$ pour laquelle $\inf f = -\infty$ bien que $x = 1/\sqrt{3}$ soit un minimiseur local.

Définition II.8. Soit $f : C \subset \mathbb{R}^N \rightarrow \mathbb{R}$, et $\bar{x} \in C$. On dit que \bar{x} est un maximiseur (resp. maximiseur local) de f sur C , s'il est un minimiseur (resp. minimiseur local) de $-f$ sur C .

Si \bar{x} est un minimiseur ou un maximiseur (local), on dit que c'est un **EXTREMA** (local).

De manière plus générale, toutes les notions et propriétés que l'on va voir par la suite porteront sur les problèmes de minimisation, et de recherche de minimiseurs, mais s'adapteront très facilement aux maximiseurs : il suffira de remplacer f par $-f$ dans les énoncés.

II.I.2 Conditions d'Optimalité du 1er ordre

Le Théorème suivant est généralement connu sous le nom de Théorème de Fermat :

Théorème II.9.

On suppose que f est différentiable en un minimiseur local \bar{x} . Alors $\nabla f(\bar{x}) = 0$.

Dans le cas où on est en présence d'une contrainte, et que le point que l'on considère est à l'intérieur de la contrainte, on obtient le même résultat :

Théorème II.10 (Théorème de Fermat : Condition Nécessaire d'Optimalité du 1er ordre).

On suppose que f est différentiable en un minimiseur local \bar{x} sur C , et que $\bar{x} \in \text{int } C$. Alors :

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

Remarque II.11. Le Théorème II.10 est encore vrai si on remplace « minimiseur local » par « maximiseur local ». Pour s'en convaincre, il suffit de remplacer f par $-f$ dans l'énoncé.

Démonstration. f admet un minimiseur local en \bar{x} , donc il existe $R > 0$ t.q.

$$(\forall x \in C \cap \mathbb{B}(\bar{x}; R)) \quad f(x) \geq f(\bar{x}). \quad (\text{II.1})$$

Comme $\bar{x} \in \text{int } C$, quitte à réduire le rayon R , on peut supposer que $\mathbb{B}_R(\bar{x}) \subset C$. Puisque f est différentiable en \bar{x} , elle admet une dérivée directionnelle en \bar{x} dans toute direction $d \in \mathbb{R}^N$, et :

$$\langle \nabla f(\bar{x}), d \rangle = \lim_{t \rightarrow 0} \frac{f(\bar{x} + td) - f(\bar{x})}{t} \geq 0,$$

où l'inégalité vient du fait que, lorsque $\|d\| |t| < R$, on a $\bar{x} + td \in \mathbb{B}(\bar{x}; R) \subset C$ et donc on peut utiliser (II.1). On a donc montré que

$$(\forall d \in \mathbb{R}^N) \quad \langle \nabla f(\bar{x}), d \rangle \geq 0,$$

Ce qui implique que $\nabla f(\bar{x}) = 0$. ■

Remarque II.12. Le résultat n'est plus valide lorsque \bar{x} n'est pas à l'intérieur de la contrainte. Un contre-exemple simple est $f(x) = x^2$, avec $C = [1, 2]$. Dans ce cas $\bar{x} = 1$ est un minimiseur global sur C , mais $f'(x) = 2 \neq 0$.

Remarque II.13. La réciproque est fausse en général, prendre par exemple $f(x) = x^3$, $f(x) = -x^2$ ou $f(x, y) = x^2 - y^2$. C'est pour cela que l'on parle de condition NÉCESSAIRE du premier ordre.

Définition II.14. Un point x où f est différentiable et $\nabla f(x) = 0$ est appelé **POINT CRITIQUE (DU PREMIER ORDRE)**. On note $\text{crit}(f)$ l'ensemble des points critiques de f .

Remarque II.15 (Minimiseurs, maximiseurs, et points selle). Si x est un point critique de f , que peut-on en dire ? Le Théorème de Fermat II.10 nous dit que tout les minimiseurs locaux et maximiseurs locaux sont des points critiques. Donc x peut être un minimiseur local ou un maximiseur local. Mais il est également possible que x ne soit ni minimiseur ni maximiseur local de f , c'est-à-dire qu'il vérifie :

pour tout voisinage V de x , il existe $x^- \in V, x^+ \in V$ tels que $f(x^-) < f(x) < f(x^+)$,

ce que l'on peut écrire de façon équivalente :

$$\exists (x_n^+)_n \in \mathbb{N}, (x_n^-)_n \in \mathbb{N} \text{ t.q. } \lim_{n \rightarrow +\infty} x_n^+ = \lim_{n \rightarrow +\infty} x_n^- = x \quad \text{et} \quad f(x_n^-) < f(x) < f(x_n^+).$$

Un tel point est appelé un **point selle**. Voir la Remarque II.13 pour des exemples de points selle.

II.I.3 Conditions d'Optimalité du 2e ordre

Théorème II.16 (Condition Nécessaire d'Optimalité, 2e ordre).

On suppose que f est deux fois différentiable en un minimiseur local \bar{x} sur C , et que $\bar{x} \in \text{int } C$. Alors

$$\nabla f(\bar{x}) = 0 \text{ et } \nabla^2 f(\bar{x}) \succeq 0.$$

Définition II.17. Un point x où f est deux fois différentiable et tel que $\nabla f(x) = 0$ et $\nabla^2 f(x) \succeq 0$ est un **POINT CRITIQUE DU DEUXIÈME ORDRE**.

Démonstration. Avant de commencer, on note $\mathbb{B}(\bar{x}, R)$ le voisinage sur lequel \bar{x} est un minimiseur local. Quitte à prendre R plus petit, on peut supposer que $\mathbb{B}(\bar{x}, R) \subset C$, puisque $\bar{x} \in \text{int } C$. On sait d'après le Théorème II.10 que $\nabla f(\bar{x}) = 0$, on ne doit donc vérifier ici que $\nabla^2 f(\bar{x}) \succeq 0$. Nous allons raisonner par l'absurde, et supposer qu'il existe $d \in \mathbb{R}^N$ tel que

$$\langle \nabla^2 f(\bar{x})d, d \rangle < 0.$$

Quitte à diviser cette inégalité par $\|d\|$, on peut supposer que $\|d\| = 1$. Dans la suite, on notera $\lambda := \langle \nabla^2 f(\bar{x})d, d \rangle < 0$.

D'après la formule de Taylor (Proposition I.78 avec $h = td$), et le fait que $\nabla f(\bar{x}) = 0$, on peut écrire, pour tout $t > 0$:

$$\begin{aligned} f(\bar{x} + td) - f(\bar{x}) &= \langle \nabla f(\bar{x}), td \rangle + \frac{1}{2} \langle \nabla^2 f(\bar{x})td, td \rangle + o(\|td\|^2) \\ &= \frac{1}{2} \langle \nabla^2 f(\bar{x})d, d \rangle t^2 + o(t^2) \\ &= \frac{\lambda}{2} t^2 + t^2 \varepsilon(t), \end{aligned}$$

où $\varepsilon(s)$ est une fonction telle que $\lim_{s \rightarrow 0} \varepsilon(s) = 0$. Maintenant, on se donne $\bar{t} < R$ tel que $\varepsilon(\bar{t}) \leq -\lambda/4$. On en déduit :

$$f(\bar{x} + \bar{t}d) - f(\bar{x}) \leq \bar{t}^2 \lambda / 4 < 0.$$

On a donc trouvé $x := \bar{x} + \bar{t}d \in C \cap \mathbb{B}(\bar{x}, R)$ tel que $f(x) < f(\bar{x})$, ce qui est une contradiction avec le fait que \bar{x} soit un minimiseur local. ■

Exemple II.18 (Réciproque). Le Théorème II.16 dit que si \bar{x} est un minimiseur local alors c'est un point critique du deuxième ordre. Est-ce que la réciproque est vraie ?

- Si on prend le cas d'une fonction quadratique (cf. Exemple II.22), on a pour tout $x \in \mathbb{R}^N$ que $\nabla^2 f(x) = A$ et $\nabla f(x) = Ax$. Donc tout point critique du second ordre est un minimiseur global. Dans ce cas la réciproque est vraie.
- Si $f(x) = x^3$, ou $-x^4$, en zéro on a $f'(0) = f''(0) = 0$ (c'est donc un point critique du deuxième ordre, au sens de la Définition II.17), mais pour autant 0 n'est pas un minimiseur local.

En général il est impossible, sans faire plus d'hypothèses, de caractériser entièrement les minimiseurs locaux avec des conditions faisant intervenir les dérivées supérieures. Mais il est possible de faire une hypothèse un peu plus forte, qui *implique* qu'un point est un minimiseur local. En gros, il faut regarder la dérivée seconde autour de x pour savoir si la fonction est localement convexe.

Théorème II.19 (Condition Suffisante d'Optimalité du 2e Ordre).

Soit f une fonction deux fois différentiable en $\bar{x} \in \text{int } C$. Supposons que

$$\nabla f(\bar{x}) = 0 \quad \text{et} \quad \nabla^2 f(\bar{x}) \succ 0.$$

Alors \bar{x} est un minimiseur local de f .

Démonstration.

Soit $\lambda = \lambda_{\min}(\nabla^2 f(\bar{x})) > 0$. D'après la formule de Taylor (Proposition I.78) (sachant que $\nabla f(\bar{x}) = 0$), il existe une fonction $\varepsilon : \mathbb{R} \rightarrow \mathbb{R}$ t.q. $\lim_{s \rightarrow 0} \varepsilon(s) = 0$ et

$$\begin{aligned} (\forall d \in \mathbb{R}^N) \quad f(\bar{x} + d) - f(\bar{x}) &= \frac{1}{2} \langle \nabla^2 f(\bar{x})d, d \rangle + \|d\|^2 \varepsilon(\|d\|) \\ &\geq \frac{\lambda}{2} \|d\|^2 + \|d\|^2 \varepsilon(\|d\|). \end{aligned}$$

Par définition de ε , il existe un $R > 0$ tel que pour tout $s \in]0, R[$, $|\varepsilon(s)| \leq \lambda/2$. Si on prend $x \in \mathbb{B}(\bar{x}; R)$ quelconque, on a $x = \bar{x} + d$ avec $d = x - \bar{x}$ et $\|d\| \leq R$, donc on déduit de ce qui précède que $\varepsilon(\|d\|) \geq -\lambda/2$, et donc que $f(x) - f(\bar{x}) \geq 0$. Ceci prouve que \bar{x} est un minimiseur local de f . ■

Remarque II.20 (Minimiseur local vs. global). Supposons que l'on ait trouvé un point \bar{x} satisfaisant aux conditions suffisantes d'optimalité du 2e ordre : le Théorème II.19 nous garantit que \bar{x} est un minimiseur local. Comment savoir s'il n'est que local, ou en fait global ?

Une bonne approche consiste à calculer $f(\bar{x})$, et à se demander si c'est le minimum de f . Il y a alors deux possibilités :

- Ou bien $f(\bar{x}) = \inf f$, auquel cas \bar{x} est bien un minimiseur global de f ,
- ou bien $f(\bar{x}) > \inf f$, ce qui implique alors que \bar{x} n'est pas un minimiseur global.

Ce deuxième cas est le plus « facile » à vérifier : il suffit en effet de réussir à trouver n'importe quel vecteur x en lequel la fonction prend une valeur plus *petite* qu'en \bar{x} : $f(x) < f(\bar{x})$.

Exercice II.21. Soit $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ définie par $f(x, y) = x^2 + y^2 \cos x$. Calculer le gradient et la matrice hessienne de f en tout point $(x, y) \in \mathbb{R}^2$. Que pouvez-vous dire du point $(0, 0)$?

Exercice II.22 (Fonction quadratique et minimiseurs). Soit $f(x) = \frac{1}{2}\langle Ax, x \rangle$, où A est une matrice symétrique. Montrer que f admet un minimiseur en 0 si et seulement si $A \succeq 0$. Est-ce que dans ce cas le minimiseur est unique ?

Exercice II.23 (Points critiques, extrema locaux et globaux). Pour les fonctions suivantes, trouver leurs points critiques et dire si ce sont des extrema locaux (ou globaux) :

- 1) $f(x, y) = x^3 + y^4$
- 2) $f(x) = (1 - x^2)^2$
- 3) $f(x, y) = x^2 + y^2 - xy^2$
- 4) $f(x) = \ln(1 + \cos x)$

Les théorèmes II.16 et II.19 nous fournissent des conditions d'optimalité vis-à-vis des minimiseurs locaux de f . On en déduit immédiatement le corollaire suivant, qui porte sur les maximiseurs locaux et les points selle :

Corollaire II.24 (CNO et CSO du 2e ordre - Maximiseurs et points selle). Soit f une fonction deux fois différentiable en $\bar{x} \in \text{int } C$.

- 1) Si $\nabla f(\bar{x}) = 0$ et $\nabla^2 f(\bar{x}) < 0$, alors \bar{x} est un maximiseur local de f sur C .
- 2) Si \bar{x} est un maximiseur local de f sur C , alors $\nabla f(\bar{x}) = 0$ et $\nabla^2 f(\bar{x}) \leq 0$.
- 3) Si $\nabla f(\bar{x}) = 0$ et $\nabla^2 f(\bar{x})$ n'est ni semi-définie positive, ni semi-définie négative, alors \bar{x} est un point selle de f .

Démonstration. 1) et 2) s'obtiennent avec les Théorèmes II.16 et II.19, en remplaçant f par $-f$. Pour le 3), si $\nabla^2 f(\bar{x})$ n'est pas semi-définie positive alors \bar{x} ne vérifie pas la CNO du

2e ordre, donc n'est pas un minimiseur local d'après II.16. De même, si $\nabla^2 f(\bar{x})$ n'est pas semi-définie négative alors \bar{x} n'est pas un minimiseur local d'après le point 2). C'est donc un point selle. ■

II.II Coercivité et existence de minimiseurs

II.II.1 Coercivité

Dans l'Exemple II.3, on voit qu'une obstruction typique à l'existence de minimiseurs est le fait que la fonction s'aplatisse indéfiniment vers l'infini, en n'atteignant jamais son infimum. Afin d'avoir un résultat d'existence, on va donc faire l'hypothèse que cela n'arrive pas. Il y a essentiellement deux manières d'y parvenir :

- 1) dire que la fonction « tend vers l'infini à l'infini » (du coup elle ne peut pas s'aplatir),
- 2) dire que la contrainte C est bornée (du coup les valeurs ne peuvent pas « tendre » vers quelque chose).

On peut combiner ces deux approches en disant simplement que la fonction « tend vers l'infini sur C » :

Définition II.25. Soit $f : C \subset \mathbb{R}^N \rightarrow \mathbb{R}$. On dit que f est **COERCIVE** sur C si

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

ce qui est une manière condensée de dire que

$$\forall (x_n)_{n \in \mathbb{N}} \subset C, \quad \lim_{n \rightarrow +\infty} \|x_n\| = +\infty \Rightarrow \lim_{n \rightarrow +\infty} f(x_n) = +\infty.$$

Lorsque $C = \mathbb{R}^N$, on dira simplement que f est coercive.

Exemple II.26. $f(x) = |x|^p$ pour $p \geq 1$ est coercive.

Exemple II.27. $f(x) = e^x$ n'est pas coercive. Par contre elle est coercive sur $[0, +\infty[$.

Exemple II.28. $f(x, y) = x^2$ n'est pas coercive, car elle est constante lorsque on fixe x .

Exercice II.29 (Coercivité). Dire à propos des fonctions suivantes si elles sont coercives.

- 1) $f(x) = (1 - x^2)^2$.
- 2) $f(x, y) = x^3 + 2y^2$.
- 3) $f(x, y) = (x - y)^2$.
- 4) $f(x, y) = \frac{x^2}{y}$ définie sur $\mathbb{R} \times]0, +\infty[$.

Remarque II.30 (Coercivité en pratique). On suppose ici $C = \mathbb{R}^N$ pour simplifier. Déterminer si une fonction est coercive ou non n'est pas une tâche facile : elle ne se ramène pas (toujours) à un simple calcul à faire, automatique. Cela demande un peu de flair, et de bien comprendre à quoi ressemble la fonction à laquelle on a affaire. Voici quelques approches :

- Votre fonction est une fonction univariée $f : \mathbb{R} \rightarrow \mathbb{R}$. Dans ce cas c'est facile, car la coercivité est équivalente à

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

Il suffit donc de calculer ces deux limites.

- Votre fonction est multivariée, et vous pensez qu'elle n'est pas coercive. Là encore c'est un cas facile, car il suffit dans ce cas de contredire la Définition II.25, et de trouver une suite $(x_n)_{n \in \mathbb{N}} \subset C$ qui vérifie :
 - $\lim \|x_n\| = +\infty$,
 - $\lim f(x_n) \neq +\infty$.
- Votre fonction est multivariée, et vous pensez qu'elle est coercive. C'est un cas un peu plus difficile, puisqu'il faut montrer que $\lim f(x_n) = +\infty$ pour **toute** suite divergente. Il serait tentant de penser que la coercivité équivaut à « fixer toutes les variables sauf une que l'on fait tendre vers $\pm\infty$ » :

$$(\forall x \in \mathbb{R}^N)(\forall i = 1..N) \quad \lim_{x_i \rightarrow \pm\infty} f(x_1, \dots, x_i, \dots, x_N) = +\infty.$$

Or ceci est **faux**. L'exercice suivant en fournit un contre-exemple.

Dans ce cas, la stratégie la plus simple est d'arriver à montrer que $f(x) \geq g(x)$, où $g(x)$ est clairement coercive. Par exemple, trouver une fonction g de la forme $g(x) = \phi(\|x\|)$, où $\phi : \mathbb{R} \rightarrow \mathbb{R}$. Dans ce cas on sait facilement montrer que ϕ est coercive, et on en déduit immédiatement que f l'est aussi.

Exercice II.31. Soit $f(x, y) = \frac{x}{y} + \frac{y}{x}$ définie sur $U =]0, +\infty[^2$.

- 1) Vérifier que, pour tout $y > 0$, $\lim_{x \rightarrow +\infty} f(x, y) = +\infty$.
- 2) Vérifier que, pour tout $x > 0$, $\lim_{y \rightarrow +\infty} f(x, y) = +\infty$.
- 3) Montrer que f n'est pas coercive sur U .

L'exercice suivant est important, et il est bon de connaître et comprendre les résultats qu'il contient :

Exercice II.32 (Fonction quadratique et coercivité).

- 1) Soient $A \in \mathcal{M}_N(\mathbb{R})$ symétrique, $b \in \mathbb{R}^N$, $c \in \mathbb{R}$. Montrer que la fonction quadratique $f(x) = \frac{1}{2}\langle Ax, x \rangle + \langle b, x \rangle + c$ est coercive si et seulement si A est définie positive.
- 2) Soient $\Phi \in \mathcal{M}_{M,N}(\mathbb{R})$, $y \in \mathbb{R}^M$. Montrer que le moindre carré $f(x) = \frac{1}{2}\|\Phi x - y\|^2$ est coercif si et seulement si Φ est injective.

On conclut cette partie avec une proposition importante, qui dit qu'une fonction est toujours coercive sur un borné.

Proposition II.33. *Soit $f : C \subset \mathbb{R}^N \rightarrow \mathbb{R}$. Si C est borné alors f est coercive sur C .*

Démonstration. C'est en fait une conséquence directe de la Définition II.25, et du fait qu'une implication $A \Rightarrow B$ est toujours vraie lorsque la proposition A est fausse. En effet, si C est bornée, il est impossible pour une suite $(x_n)_{n \in \mathbb{N}} \subset C$ de vérifier $\lim_{n \rightarrow +\infty} \|x_n\| = +\infty$. ■

Le lien entre « coercivité » et « borné » n'est d'ailleurs pas anodin ! En effet, la Proposition suivante montre que la coercivité d'une fonction f peut entièrement être caractérisée par le fait que ses sous-niveaux soient bornés.

Proposition II.34 (Coercivité et sous-niveaux bornés). *Soient $f : U \subset \mathbb{R}^N \rightarrow \mathbb{R}$, $C \subset U$, et notons, pour tout $r \in \mathbb{R}$, le sous-niveau de f*

$$[f \leq r] := \{x \in U \mid f(x) \leq r\}.$$

Alors f est coercive sur C si et seulement si $C \cap [f \leq r]$ est borné pour tout $r \in \mathbb{R}$.

Démonstration. Dans cette preuve on notera $\Omega_r := C \cap [f \leq r]$.

\Rightarrow : Supposons que f soit coercive sur C , donnons-nous $r \in \mathbb{R}$ quelconque, et montrons que Ω_r est borné. Pour cela, raisonnons par l'absurde et supposons que Ω_r ne soit pas borné. Alors il doit exister une suite $(x_n)_{n \in \mathbb{N}} \subset \Omega_r$ telle que $\|x_n\| \rightarrow +\infty$. On a donc une suite qui diverge, contenue dans C : notre hypothèse (f coercive) nous permet donc de déduire que $f(x_n)$ tend vers $+\infty$. En particulier, cela veut dire qu'à partir d'un certain rang, $f(x_n) > r$, ce qui contredit $x_n \in \Omega_r$. L'implication est donc démontrée.

\Leftarrow : Supposons que Ω_r soit borné pour tout $r \in \mathbb{R}$, et montrons que f est coercive sur C . Supposons donc qu'il existe une suite $(x_n)_{n \in \mathbb{N}} \subset C$ telle que $\|x_n\| \rightarrow +\infty$, et montrons que $f(x_n)$ tend vers $+\infty$. Fixons pour cela un $r \in \mathbb{R}$ quelconque. Puisque la suite x_n diverge, et que Ω_r est borné par hypothèse, cela veut dire qu'à partir d'un certain rang, $x_n \notin \Omega_r$. Or $\Omega_r = C \cap [f \leq r]$, et on sait que $x_n \in C$. Donc cela veut dire qu'à partir d'un certain rang, $x_n \notin [f \leq r]$. Autrement dit, que $f(x_n) > r$. Ceci étant vrai pour tout $r \in \mathbb{R}$, on conclut que $f(x_n)$ tend vers $+\infty$. ■

II.II.2 Existence de minimiseurs

Théorème II.35 (Existence si continue coercive). *Soit $f : C \subset \mathbb{R}^N \rightarrow \mathbb{R}$. Supposons que :*

- a) C est fermé,
- b) f est continue en tout point de C ,
- c) f est coercive sur C .

Alors f admet un minimiseur global sur C .

Remarque II.36 (Pas de réciproque). La réciproque de ce Théorème est évidemment fausse : l'existence d'un minimiseur global n'implique pas la coercivité. Par exemple, $f(x, y) = x^2$ ou $f(x, y) = 18$ ne sont pas coercives mais admettent des minimiseurs globaux.

Pour prouver ce résultat on aura besoin d'un Lemme élémentaire sur l'existence de suites minimisantes :

Lemme II.37 (Suite minimisante). Pour tout ensemble $C \subset \mathbb{R}^N$ et toute fonction $f : C \rightarrow \mathbb{R}$, il existe une suite $(x_n)_{n \in \mathbb{N}} \subset C$ telle que $\lim_{n \rightarrow +\infty} f(x_n) = \inf_C f$.

Démonstration. On introduit l'ensemble $V := \{f(x), x \in C\} \subset \mathbb{R}$, qui vérifie par définition que $\inf V = \inf_C f$. Distinguons deux cas de figure :

- **Cas $\inf V \in \mathbb{R}$.** Par définition de l'infimum d'une partie de \mathbb{R} , on a que, pour tout $\varepsilon > 0$, il existe $v_\varepsilon \in V$ tel que

$$\inf V \leq v_\varepsilon < \inf V + \varepsilon.$$

Or, par définition de V , il existe un $x_\varepsilon \in C$ tel que $v_\varepsilon = f(x_\varepsilon)$. Ainsi, on a que pour tout $\varepsilon > 0$, il existe $x_\varepsilon \in C$ tel que

$$\inf_C f \leq f(x_\varepsilon) < \inf_C f + \varepsilon.$$

En prenant $\varepsilon = 1/n$ et en passant à la limite, on obtient que $\lim_{n \rightarrow \infty} f(x_n) = \inf_C f$.

- **Cas $\inf V = -\infty$.** Dans ce cas, pour tout $n \in \mathbb{N}$ il existe un point qu'on note $x_n \in C$ tel que $f(x_n) < -n$. On en déduit que $\lim_{n \rightarrow \infty} f(x_n) = -\infty = \inf_C f$, qui est ce que l'on voulait démontrer.



Démonstration du Théorème II.35. D'après le Lemme précédent, on peut invoquer une suite minimisante, c'est-à-dire une suite $(x_n)_{n \in \mathbb{N}} \subset C$ telle que $\lim_{n \rightarrow \infty} f(x_n) = \inf_C f$. On utilise maintenant le fait que f soit coercive : puisque $\lim_{n \rightarrow \infty} f(x_n) \neq +\infty$, la Définition II.25 nous permet de dire, par contraposée, que la propriété $\lim_{n \rightarrow \infty} \|x_n\| = +\infty$ est fausse. En d'autres termes, $(x_n)_{n \in \mathbb{N}}$ admet une sous-suite bornée. Par compacité, on en déduit que $(x_n)_{n \in \mathbb{N}}$ admet une (sous-)sous-suite convergente dans \mathbb{R}^N : on note $(x_{n_k})_{k \in \mathbb{N}}$ cette sous-suite, et \bar{x} sa limite dans \mathbb{R}^N . Comme C est fermé et $x_{n_k} \in C$, on sait que $\bar{x} \in C$. Comme f est continue sur C , on en déduit que $f(\bar{x}) = \lim_{n \rightarrow \infty} f(x_{n_k}) = \inf_C f$. Ceci prouve que \bar{x} est un minimiseur de f sur C .



Exercice II.38. Montrer que si on enlève la moindre des trois hypothèses du Théorème II.35, alors la conclusion n'est plus vraie.

Le Théorème II.35 est une version plus générale de ce résultat que vous connaissez déjà certainement :

Corollaire II.39 (Théorème des valeurs extrêmes - Bolzano, 1817). Soit $f : C \subset \mathbb{R}^N \rightarrow \mathbb{R}$. Si f est continue sur C compact, alors f admet un minimiseur global sur C .

Démonstration. C'est une conséquence immédiate du Théorème II.35 et de la Proposition II.33. ■

Exercice II.40 (Existence de minimiseurs 2). Déterminer si le problème d'optimisation sous contraintes $\inf_{x \in C} f(x)$ admet un minimiseur, pour les cas suivants :

- 1) $f(x) = \frac{1}{2}\langle Ax, x \rangle + \langle b, x \rangle$, avec $A \in \mathcal{M}_N(\mathbb{R})$ symétrique définie positive, $b \in \mathbb{R}^N$ et $C := \{x \in \mathbb{R}^N \mid (\forall i = 1, \dots, N) \quad x_i \geq c_i\}$, où $c_i \in \mathbb{R}$.
- 2) $f : \mathbb{R}^N \rightarrow \mathbb{R}$ est une fonction continue et $C := \{x \in \mathbb{R}^N \mid \sum_{i=1}^N a_i x_i^2 \leq 1 \text{ et } \sum_{i=1}^N x_i = 1\}$ (avec $a_i > 0$ fixés).
- 3) $f(x) = d(x, y)$, où $y \in \mathbb{R}^N$ est fixé, et d est la distance euclidienne sur \mathbb{R}^N ; et C fermé non vide. Comment décririez-vous les minimiseurs de f sur C ? On montrera de plus qu'il n'y a pas en général unicité du minimiseur.

Un second exercice important sur les fonctions quadratiques, qui montre que les moindres carrés $\|\Phi x - y\|^2$ admettent toujours un minimiseur global :

Exercice II.41 (Fonction quadratique et minimiseurs). Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique semi-définie positive.

- 1) Montrer que

$$(\forall x \in \text{Ker } A^\perp) \quad \langle Ax, x \rangle \geq \sigma \|x\|^2,$$

où σ est la plus petite valeur propre non nulle de A .

Indication : Toute matrice symétrique est diagonalisable dans une base orthogonale de vecteurs propres. Cette question est plus difficile que les autres, n'hésitez pas à la faire en dernier si vous bloquez.

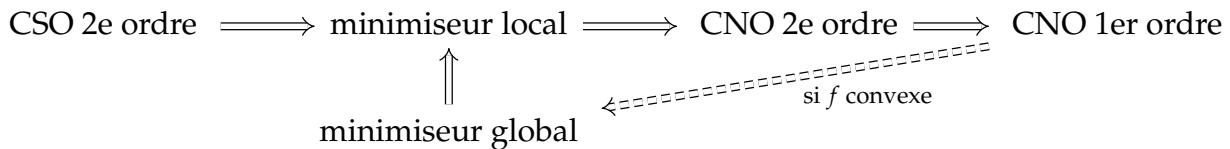
- 2) Soit $f(x) = \frac{1}{2}\langle Ax, x \rangle + \langle b, x \rangle + c$, où $b \in \mathbb{R}^N$ et $c \in \mathbb{R}$. Montrer que f admet un minimiseur sur $C = \text{Ker } A^\perp$.
- 3) Justifier que f n'admet pas nécessairement un minimiseur sur \mathbb{R}^N . A votre avis, quelle condition sur A et b faut-il pour que cela soit vrai? (on attend une conjecture plutôt qu'une preuve)
- 4) Soit $\Phi \in \mathcal{M}_{M,N}(\mathbb{R})$, $y \in \mathbb{R}^M$. Montrer que $f(x) = \frac{1}{2}\|\Phi x - y\|^2$ admet un minimiseur sur \mathbb{R}^N .

II.III Récapitulatif du Chapitre

Ici $C \subset U \subset \mathbb{R}^N$, où U est un ouvert de \mathbb{R}^N , et C est une contrainte non vide. On considère une fonction $f : U \rightarrow \mathbb{R}$, et le problème d'optimisation associé

$$\text{minimiser}_{x \in C} f(x).$$

Conditions nécessaire et suffisante d'optimalité (locale) Si $\bar{x} \in \text{int } C$, alors nous avons les implications suivantes :



- Condition Nécessaire d'Optimalité (CNO) d'ordre 1 : Si $\bar{x} \in \text{int } C$ est un minimiseur local de f sur C , alors $\nabla f(\bar{x}) = 0$.
 - La réciproque est fausse en général (par exemple $f(x) = x^3$).
 - La condition $\bar{x} \in \text{int } C$ est automatiquement vérifiée si il n'y a pas de contraintes puisque $C = \mathbb{R}^N$ est un ouvert.
 - La condition $\bar{x} \in \text{int } C$ est essentielle, le résultat est faux lorsque $\bar{x} \in \text{bd } C$.
- Condition Nécessaire d'Optimalité (CNO) d'ordre 2 : Si $\bar{x} \in \text{int } C$ est un minimiseur local de f sur C , alors $\nabla f(\bar{x}) = 0$ et $\nabla^2 f(\bar{x}) \succeq 0$.
 - La réciproque est fausse en général (par exemple $f(x) = -x^4$), il faut plus :
- Condition Suffisante d'Optimalité (CSO) d'ordre 2 : Si $\bar{x} \in \text{int } C$ vérifie $\nabla f(\bar{x}) = 0$ et $\nabla^2 f(\bar{x}) \succ 0$, alors \bar{x} est un minimiseur local de f sur C .
 - La réciproque est fausse en général (par exemple $f(x) = x^4$).

Dans le prochain chapitre, on verra que l'hypothèse clé pour obtenir des réciproques à ces résultats et de supposer que le problème est *convexe*.

Existence de minimiseurs (globaux)

- Si f est coercive sur C , alors f admet au moins un minimiseur global sur C .
- Si C est borné, alors f est coercive sur C .

Chapitre III

Optimisation convexe

III.I Convexité et globalité des minimiseurs

III.I.1 Ensemble convexe

Définition III.1. Etant donné deux points x, y dans \mathbb{R}^N , on définit l'intervalle qui les relie par

$$[x, y] := \{(1 - \alpha)x + \alpha y \mid \alpha \in [0, 1]\}.$$

Définition III.2. Soit $C \subset \mathbb{R}^N$. On dit que l'ensemble C est **CONVEXE** si

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

Autrement dit, il faut et il suffit que pour toute paire de points x, y dans C , l'intervalle $[x, y]$ qui relie ces points soit également contenu dans C (cf. Figure III.1).

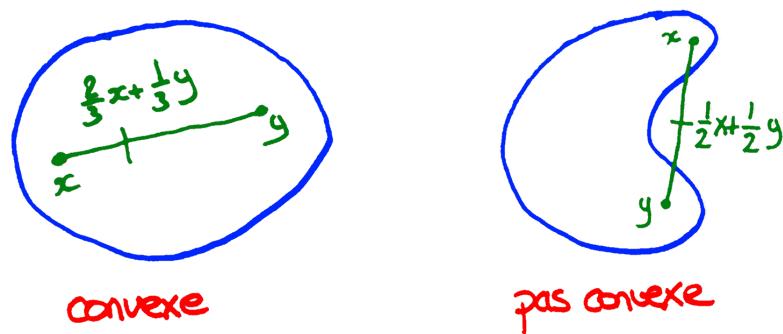


FIGURE III.1 – Convexité d'un ensemble

Exemple III.3. La boule unité $\mathbb{B}(0, 1) = \{x \in \mathbb{R}^N \mid \|x\| \leq 1\}$ est convexe. La sphère unité $\mathbb{S}(0, 1) = \{x \in \mathbb{R}^N \mid \|x\| = 1\}$, elle, n'est pas convexe car elle est creuse.

Exemple III.4. Les sous-ensembles convexes de \mathbb{R} sont les intervalles.

Exercice III.5 (Convexité et intersection). Soit $\mathcal{N} : \mathbb{R}^N \rightarrow \mathbb{R}_+$ une norme quelconque. Montrer que la boule unité (fermée) pour cette norme est nécessairement convexe.

Exercice III.6. Montrer que l'intersection de deux ensembles convexes est encore convexe. En déduire que l'intersection d'un nombre fini d'ensemble convexes est convexe.

III.I.2 Fonction convexe

Définition III.7. Soit $f : U \subset \mathbb{R}^N \rightarrow \mathbb{R}$, et $C \subset U$. On dit que f est **CONVEXE** sur C si C est convexe et que

$$\forall \alpha \in [0, 1], \forall (x, y) \in C^2, \quad f((1 - \alpha)x + \alpha y) \leq (1 - \alpha)f(x) + \alpha f(y).$$

On notera¹ $\Gamma_0(C)$ l'ensemble des fonctions convexes sur C . Si $C = \mathbb{R}^N$ on dira simplement que f est convexe.

Proposition III.8. Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$. Alors ces deux propriétés sont équivalentes :

- 1) f est convexe,
- 2) l'épigraphhe² de f est convexe, ce dernier étant défini par :

$$\text{epi } f = \{(x, y) \in \mathbb{R}^N \times \mathbb{R} \mid f(x) \leq y\} \subset \mathbb{R}^N \times \mathbb{R}.$$

Démonstration. Voir TD. ■

On peut donner une caractérisation géométrique similaire pour la convexité d'une fonction sur une contrainte :

Proposition III.9. Soit $f : U \subset \mathbb{R}^N \rightarrow \mathbb{R}$, et $C \subset U$. Alors ces deux propriétés sont équivalentes :

- 1) f est convexe sur C ,

¹Il est assez difficile de retrouver d'où vient la notation Γ_0 . Néanmoins il semblerait que cela remonte aux premiers travaux de Fenchel (1951) et Moreau (1965), dans lesquels Γ_0 décrit l'ensemble des fonctions convexes semi-continues inférieurement et propres (pas constantes à l'infini). Le choix d'utiliser la lettre Γ semblerait être en dualité avec la lettre C (pour convexe), Γ étant également la troisième lettre de l'alphabet grec. Quand à l'indice 0 son sens s'est perdu mais dans ce cours on va lui donner un signification (cf. Section sur les fonctions fortment convexes). Une discussion intéressante à ce sujet ici <https://mathoverflow.net/questions/262851/why-are-gamma-0-functions-called-this/262861>

²« epi » est un préfixe qui veut dire « au-dessus ». C'est l'opposé de « hypo » qui nous est plus familier.

2) l'épigraphe de f sur C est convexe, ce dernier étant défini par :

$$\text{epi}_C f = \{(x, y) \in \mathbb{R}^N \times \mathbb{R} \mid x \in C, f(x) \leq y\} \subset \mathbb{R}^N \times \mathbb{R}.$$

Démonstration. Voir TD. ■

Proposition III.10. Soient $f, g : U \subset \mathbb{R}^N \rightarrow \mathbb{R}$, et $C \subset U$. Si f et g sont convexes sur C , alors $f + g$ est convexe sur C .

Démonstration. Cf. TD. ■

Proposition III.11. Soit $f : \mathbb{R}^M \rightarrow \mathbb{R}$ une fonction convexe, et $A \in \mathcal{M}_{N,M}(\mathbb{R})$. Alors $f \circ A$ est convexe.

Démonstration. Cf. TD. ■

Proposition III.12. Soit $f : U \subset \mathbb{R}^N \rightarrow \mathbb{R}$ et $C \subset U$. Si f est convexe sur C , alors $\text{argmin}_C f$ est un ensemble convexe.

Démonstration. Cf. TD. ■

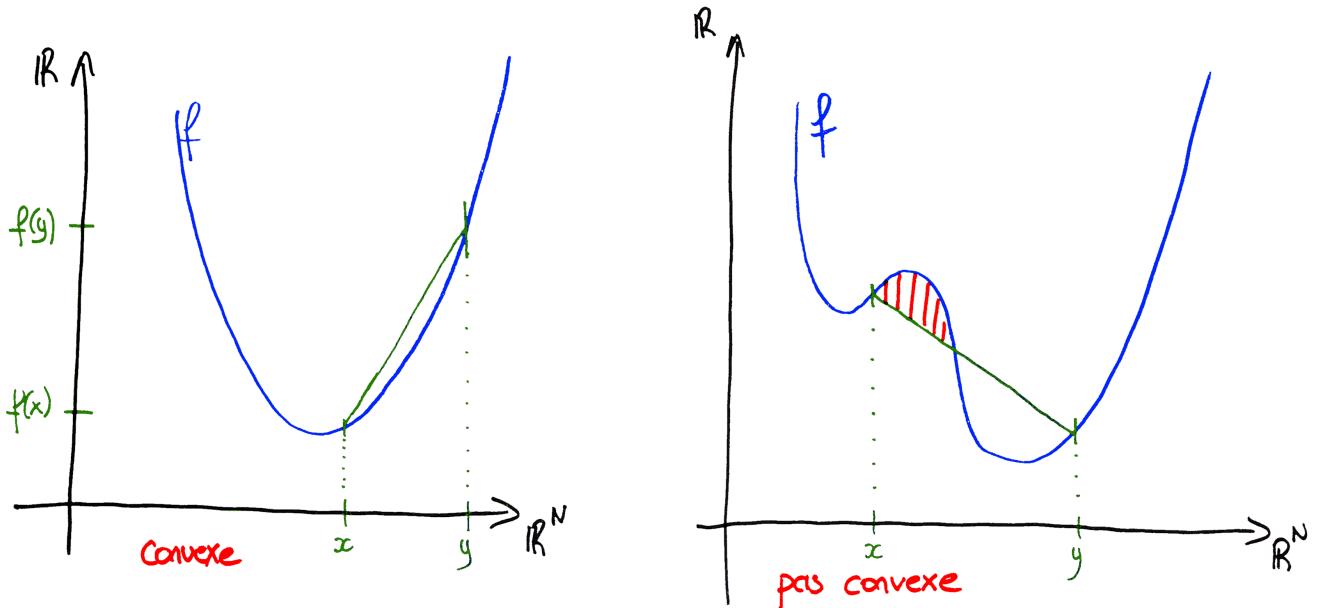


FIGURE III.2 – Convexité d'une fonction

III.I.3 Caractérisation de la convexité pour les fonctions univariées

On s'intéresse d'abord ici aux fonctions d'une seule variable. Dans cette sous-section III.I.3, on supposera toujours que $U \subset \mathbb{R}$ est un ouvert, et que $I \subset U$ est un intervalle.

Proposition III.13 (Convexité via dérivée). Soient $f : U \subset \mathbb{R} \rightarrow \mathbb{R}$ une fonction dérivable, et $I \subset U$ un intervalle. Les propriétés suivantes sont alors équivalentes :

- i) f est convexe sur I , c-à-d $f \in \Gamma_0(I)$;
- ii) $(\forall (x, y) \in I^2) \quad f(y) \geq f(x) + f'(x)(y - x)$;
- iii) f' est croissante sur I .

Remarque III.14. L'équation de l'hyperplan tangent au graphe de f , au point $(x_0, f(x_0)) \in I \times \mathbb{R}$, s'écrit

$$y = f(x_0) + f'(x_0)(x - x_0), \quad \text{pour } x \in \mathbb{R}, y \in \mathbb{R}.$$

La relation ii) signifie géométriquement que le graphe de f est *au-dessus* de son hyperplan tangent en tout point (cf. Figure III.3).

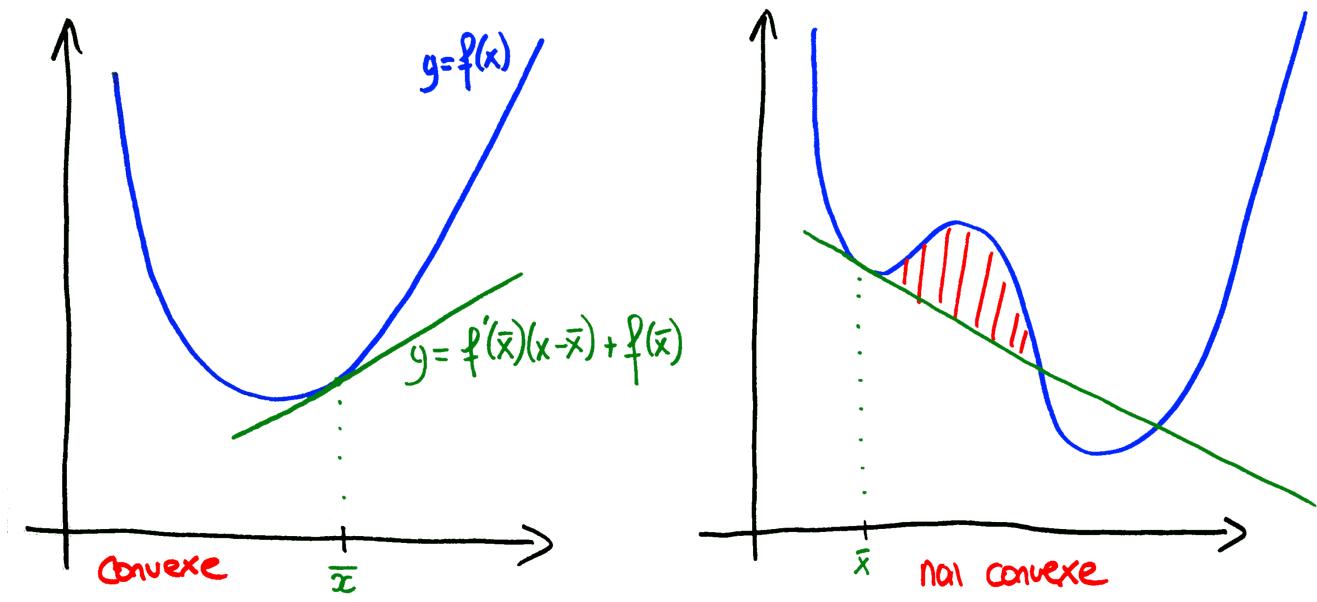


FIGURE III.3 – Convexité d'une fonction via l'hyperplan tangent

Démonstration. i) \Rightarrow ii). Soient $(x, y) \in I^2$ quelconques. Pour $\alpha \in]0, 1[$, on pose $z_\alpha := (1 - \alpha)x + \alpha y$. On a alors $f(z_\alpha) \leq (1 - \alpha)f(x) + \alpha f(y) = f(x) + \alpha(f(y) - f(x))$, donc

$$f(y) - f(x) \geq \frac{1}{\alpha}(f(z_\alpha) - f(x)) \xrightarrow{\alpha \rightarrow 0^+} f'(x)(y - x).$$

ii) \Rightarrow i) : On a

$$f(x) \geq f(z_\alpha) + f'(z_\alpha)(x - z_\alpha) \quad (\text{III.1})$$

$$f(y) \geq f(z_\alpha) + f'(z_\alpha)(y - z_\alpha). \quad (\text{III.2})$$

En sommant $(1 - \alpha)$ fois la relation (III.1) et α fois la relation (III.2), et en utilisant le fait que $(1 - \alpha)(x - z_\alpha) + \alpha(y - z_\alpha) = 0$, on obtient l'inégalité de convexité.

ii) \Rightarrow iii) : On écrit

$$\begin{aligned} f(y) &\geq f(x) + f'(x)(y - x) \\ f(x) &\geq f(y) + f'(y)(x - y). \end{aligned}$$

En sommant ces inégalités, on obtient l'inégalité désirée : $(f'(y) - f'(x))(y - x) \geq 0$.

iii) \Rightarrow ii) : Soit $g(t) := f((1 - t)x + ty)$ pour $t \in [0, 1]$. Notons que g est dérivable sur $[0, 1]$, car f est dérivable sur un ouvert U , et que x, y appartiennent à l'intervalle $I \subset U$. On calcule que $g'(t) = f'(z_t)(y - x)$, et en particulier que $g'(0) = f'(x)(y - x)$. Donc il nous suffit de montrer que $g(1) - g(0) - g'(0) \geq 0$. D'après notre hypothèse, on a

$$g'(t) - g'(0) = f'(z_t) - f'(x)(y - x) = \frac{1}{t}(f'(z_t) - f'(x))(z_t - x) \geq 0.$$

D'autre part, comme g est continue sur $[0, 1]$ et dérivable sur $]0, 1[$, on peut utiliser le théorème des accroissements finis qui nous dit qu'il existe $c \in]0, 1[$ tel que $\frac{g(1) - g(0)}{1} = g'(c)$. En combinant ces deux résultats, on en déduit que $g(1) - g(0) \geq g'(0)$, ce qui donne l'inégalité désirée. ■

Lemme III.15. Soient $f : U \subset \mathbb{R} \rightarrow \mathbb{R}$ une fonction dérivable, et $I \subset U$ un intervalle. Alors

$$f \text{ est croissante sur } I \Leftrightarrow f'(x) \geq 0 \text{ pour tout } x \in I.$$

Démonstration. Vu en Analyse L2, on rappelle la preuve ici.

\Rightarrow : Soit $x \in I$. Puisque I est un intervalle, il existe $h_n \neq 0$ tel que $h_n \rightarrow 0$ et $x + h_n \in I$. Puisque f est croissante sur I , on voit qu'on a $\frac{f(x+h_n) - f(x)}{h_n} \geq 0$ et ce quelque soit le signe de h_n . En passant à la limite, on en déduit que $f'(x) \geq 0$.

\Leftarrow : Soient $a < b$ dans I . On sait que f est dérivable sur $[a, b]$, donc on peut utiliser le Théorème des accroissements finis, qui nous fournit un $c \in]a, b[$ (en particulier $c \in I$) tel que $f(b) - f(a) = f'(c)(b - a)$. On en déduit donc que $f(a) < f(b)$. ■

Théorème III.16 (Convexité via Dérivée seconde). Soient $f : U \subset \mathbb{R} \rightarrow \mathbb{R}$ une fonction deux fois dérivable, et $I \subset U$ un intervalle. Alors les propriétés suivantes sont équivalentes :

- i) f est convexe sur I , c-à-d $f \in \Gamma_0(I)$;
- ii) $(\forall x \in I) \quad f''(x) \geq 0$.

Démonstration. Immédiat en combinant les deux résultats précédents. ■

III.I.4 Caractérisation de la convexité pour les fonctions multivariées

Afin d'étudier la convexité des fonctions multivariées à l'aide des résultats de la section précédente, on va utiliser le Lemme suivant :

Lemme III.17. Soit $f : U \subset \mathbb{R}^N \rightarrow \mathbb{R}$, et $C \subset U$ convexe. Alors f est convexe si et seulement si

$$(\forall x, y \in C) \quad \text{la fonction } g_{x,y} : t \in [0, 1] \mapsto f((1-t)x + ty) \text{ est convexe sur } [0, 1].$$

Démonstration.

\Rightarrow Soient $x, y \in C$, et montrons que $g_{x,y}$ est convexe sur $[0, 1]$. Pour cela, on se donne $t_1, t_2 \in [0, 1]$, $\alpha \in [0, 1]$, et on va montrer que

$$g((1-\alpha)t_1 + \alpha t_2) \leq (1-\alpha)g(t_1) + \alpha g(t_2).$$

Le membre de gauche peut se réécrire ainsi :

$$\begin{aligned} g((1-\alpha)t_1 + \alpha t_2) &= f([1 - (1-\alpha)t_1 - \alpha t_2]x + [(1-\alpha)t_1 + \alpha t_2]y) \\ &= f((1-\alpha)[(1-t_1)x + t_1 y] + \alpha[(1-t_2)x + t_2 y]) \end{aligned}$$

En utilisant la convexité de f en les points $(1-t_1)x + t_1 y$ et $(1-t_2)x + t_2 y$, on en conclut que

$$\begin{aligned} g((1-\alpha)t_1 + \alpha t_2) &\leq (1-\alpha)f((1-t_1)x + t_1 y) + \alpha f((1-t_2)x + t_2 y) \\ &= (1-\alpha)g(t_1) + \alpha g(t_2). \end{aligned}$$

\Leftarrow Soient $x, y \in C$ quelconques, et $\alpha \in [0, 1]$. On peut alors utiliser la convexité de $g_{x,y}$ pour écrire

$$\begin{aligned} f((1-\alpha)x + \alpha y) &= g_{x,y}(\alpha) = g_{x,y}((1-\alpha).0 + \alpha.1) \\ &\leq (1-\alpha)g_{x,y}(0) + \alpha g_{x,y}(1) \\ &= (1-\alpha)f(x) + \alpha f(y). \end{aligned}$$
■

Proposition III.18 (Convexité via le gradient). Soit $f : U \subset \mathbb{R}^N \rightarrow \mathbb{R}$, différentiable sur U , et $C \subset U$ convexe. Alors f est convexe si et seulement si

$$(\forall x, y \in C) \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle. \quad (\text{III.3})$$

Démonstration. On va réutiliser ici les notations du Lemme III.17, et son résultat.

\Rightarrow : Supposons que f soit convexe et prouvons (III.3). Soient donc $x, y \in C$, on sait alors via le Lemme III.17 que $g := g_{x,y} : [0, 1] \rightarrow \mathbb{R}$ est convexe. Par ailleurs, puisque $x, y \in C \subset U$ ouvert, il existe en fait un $\varepsilon > 0$ tel que g soit bien définie et dérivable sur $]-\varepsilon, 1 + \varepsilon[$, avec $g'(t) = \langle \nabla f((1-t)x + ty), y - x \rangle$. On peut donc appliquer la Proposition III.13 qui nous dit que

$$(\forall a, b \in [0, 1]) \quad g(b) \geq g(a) + g'(a)(b - a).$$

On voit qu'en prenant $b = 1$ et $a = 0$, on obtient bien

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

\Leftarrow : Supposons (III.3) et prouvons que f est convexe. Via le Lemme III.17, il suffit donc de fixer $x, y \in C$ et de montrer que $g := g_{x,y} : [0, 1] \rightarrow \mathbb{R}$ est convexe sur $[0, 1]$. Donc, via la Proposition III.13, il suffit de montrer que

$$(\forall a, b \in [0, 1]) \quad g(b) \geq g(a) + g'(a)(b - a),$$

ce qui se réécrit par définition de g

$$f((1-b)x + by) \geq f((1-a)x + ay) + \langle \nabla f((1-a)x + ay), y - x \rangle (b - a).$$

Or cette inégalité est exactement ce que l'on obtient lorsque dans (III.3) on remplace y par $(1-b)x + by$ et x par $(1-a)x + ay$. ■

Théorème III.19 (Convexité via Hessienne). Soient $f : U \subset \mathbb{R}^N \rightarrow \mathbb{R}$, deux fois différentiable sur U , et $C \subset U$ convexe. Considérons les propriétés suivantes :

- i) $(\forall x \in C) \quad \nabla^2 f(x) \succeq 0$;
- ii) f est convexe sur C , c-à-d $f \in \Gamma_0(C)$.

Alors i) \Rightarrow ii), et l'équivalence i) \Leftrightarrow ii) est vraie si C est ouvert.

Démonstration.

i) \Rightarrow ii). Afin de montrer que f est convexe sur C , nous allons montrer que $g_{x,y}$ est convexe pour tout $x, y \in C$, puis conclure avec le Lemme III.17 précédent. D'après le Théorème III.16, il nous suffit de montrer que $g''_{x,y}$ est positive, où

$$g''_{x,y}(t) = \langle \nabla^2 f(x + t(y-x))(y-x), y-x \rangle. \quad (\text{III.4})$$

Or notre hypothèse, combinée avec (III.4), et le fait que C est convexe, impliquent que c'est bien le cas.

ii) \Rightarrow i). Soit $x \in C$. Afin de montrer que $\nabla^2 f(x) \succeq 0$, on va prendre $d \in \mathbb{R}^N$ quelconque, et montrer que $\langle \nabla^2 f(x)d, d \rangle \geq 0$. Puisque C est ouvert, il existe $\delta > 0$ tel que $B(x, \delta) \subset C$. Donc $y := x + \varepsilon d$ appartient à C pour $0 < \varepsilon < \delta/2\|d\|$. On peut donc faire appel à la

fonction $g_{x,y}$ qui est convexe sur $[0, 1]$ d'après le Lemme III.17. De plus sa dérivée seconde est bien définie sur $[0, 1]$ (et donnée par (III.4)) puisque $x, y \in C \subset U$ ouvert. En particulier, on peut utiliser le Théorème III.16, et en regardant $g''(0)$, on voit que

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

Or $y - x = \varepsilon d$, d'où le résultat. ■

Remarque III.20 (Cas $N = 1$). Pour $N = 1$, on retrouve le critère usuel : « f est convexe ssi f'' est positive ».

Remarque III.21 (Positivité d'une famille de matrices). Pour une fonction multivariée, vérifier en pratique si une fonction est convexe revient à vérifier que la matrice Hессienne est semi-définie positive. Il est donc pour cela important d'être capable de déterminer aisément si une matrice symétrique est semi-définie positive ou non (cf. Chapitre I). Il est également important de souligner qu'il faut vérifier la positivité d'une *famille* de matrices, à savoir

$$\{\nabla^2 f(x) : x \in C\}.$$

Dans le cas où C est ouvert, si une seule de ces Hessiennes échoue à être semi-définie positive, alors la fonction ne sera pas convexe.

Remarque III.22 (Convexité sur une contrainte non ouverte). Si $f \in \Gamma_0(C)$, que peut-on dire de $\nabla^2 f(x)$ pour $x \in C$?

- Lorsque C est ouvert, le Théorème III.19 nous garantit que $\nabla^2 f(x) \succeq 0$.
- Lorsque $\text{int } C \neq \emptyset$ et $f \in C^2(U)$, alors on peut également conclure que $\nabla^2 f(x) \succeq 0$. En effet on sait que la Hessianne est semi-définie positive sur $\text{int } C$, en appliquant le Théorème III.19 à $\text{int } C$, qui est ouvert. De plus, on suppose que $\nabla^2 f$ est continue, donc les valeurs propres de $\nabla^2 f(x)$ sont continues en x . Puisque $C \subset \overline{\text{int } C}$, on déduit en passant à la limite que la Hessianne est également semi-définie positive sur le bord de C .
- Lorsque $\text{int } C = \emptyset$ on ne peut pas se prononcer. En effet sur un C d'intérieur vide on est « aveugle » par rapport à ce que fait f en dehors de C , ce qui empêche de décrire le comportement de la Hessianne dans les directions qui « pointent » vers l'extérieur. On peut par exemple considérer le contre-exemple de la fonction $f(x) = x^3$ qui est convexe (car constante !) sur $C = \{-1\}$, alors que $f''(x) = -6 < 0$ sur C . Si on veut un exemple avec une contrainte qui ne soit pas un singleton, on peut également considérer $f(x, y) = x^3$ qui est convexe sur $C = \{(x, y) \in \mathbb{R}^2 \mid x = -1\}$. On reverra ce genre de problème lorsqu'on étudiera en détail les problèmes d'optimisation sous contraintes dans le Chapitre V (voir en particulier la Remarque V.46).

Proposition III.23. Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ une fonction quadratique : $f(x) = \langle Ax, x \rangle + \langle b, x \rangle + c$. Alors f est convexe si et seulement si $A \in \mathcal{M}_N(\mathbb{R})$ est semi-définie positive.

Démonstration. Cf. TD. ■

III.I.5 Convexité et minimiseurs

Lorsqu'une fonction est convexe, elle ressemble à un U , et donc elle n'a pas de minimiseur locaux, mais que des minimiseur globaux. Cela provient du fait que la notion de convexité est une notion *globale*; par exemple il faut que la Hessienne soit semi-définie positive en tout point.

Théorème III.24 (Convexe : Minimum local = global). Soit $C \subset \mathbb{R}^N$ convexe et $f \in \Gamma_0(C)$. Soit $\bar{x} \in C$ un minimiseur local de f sur C . Alors \bar{x} est un minimiseur global de f sur C .

Démonstration. Soit $R > 0$ tel que \bar{x} soit un minimiseur de f sur $C \cap \mathbb{B}(\bar{x}, R)$. Soit $x \in C$ quelconque, et montrons que $f(\bar{x}) \leq f(x)$. Pour simplifier on suppose $x \neq \bar{x}$. Posons $d = x - \bar{x}$. Alors $\bar{x} + td \in \mathbb{B}(\bar{x}, R)$, pourvu que $0 < t\|d\| < R$, et donc $f(\bar{x}) \leq f(\bar{x} + td)$. Or on peut écrire $\bar{x} + td = (1 - t)\bar{x} + tx$, donc par convexité on a :

$$f(\bar{x}) \leq f(\bar{x} + td) \leq (1 - t)f(\bar{x}) + tf(x),$$

que l'on peut réécrire :

$$0 \leq t(f(x) - f(\bar{x})).$$

On peut alors conclure après avoir divisé par $t > 0$. ■

Une seconde propriété très importante des fonctions convexes est que tout point critique du premier ordre est un minimiseur global. Lorsque la fonction est deux fois différentiable, c'est une conséquence directe du Théorème II.19 et Proposition III.19.i). En fait, cela reste vrai même si la fonction n'est pas deux fois différentiable.

Théorème III.25 (Convexe : Point critique = min global). Soit $C \subset \mathbb{R}^N$ convexe et $f \in \Gamma_0(C)$. Si f est différentiable en $\bar{x} \in \text{int } C$, alors $\nabla f(\bar{x}) = 0$ si et seulement si \bar{x} est un minimiseur global de f sur C .

Remarque III.26 (Gare au bord de la contrainte!). Comme on l'a dit précédemment, la réciproque est **fausse** en général lorsque \bar{x} appartient au bord de la contrainte C . On verra au chapitre V ce qu'il se passe dans ce cas.

Il faut également noter qu'il existe aussi un résultat analogue lorsque la fonction n'est pas différentiable en \bar{x} , mais c'est hors programme (cf. Cours du Master MIDS).

Démonstration. Comme $\bar{x} \in \text{int } C$, il existe $R > 0$ tel que $\mathbb{B}(\bar{x}, R) \subset C$. Pour tout $x \in \mathbb{B}(\bar{x}, R)$, on peut écrire d'après la Proposition III.18 :

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

Ceci montre donc que \bar{x} est un minimiseur local de f sur C . On conclut alors avec le Théorème III.24. ■

Proposition III.27 (Fonction quadratique et minimiseurs). Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ une fonction quadratique : $f(x) = \frac{1}{2}\langle Ax, x \rangle + \langle b, x \rangle + c$, avec $A \in \mathcal{M}_N(\mathbb{R})$, $b \in \mathbb{R}^N$ et $c \in \mathbb{R}$. Alors f admet des minimiseurs si et seulement si $A \succeq 0$ et $b \in \text{Im } A$. Dans ce cas, $\arg\min f = \{x \in \mathbb{R}^N \mid Ax + b = 0\}$.

Démonstration. Cf. TD. ■

III.II Forte convexité : existence et unicité du minimiseur

III.II.1 Fonction fortement convexe

Définition III.28. Soit $f : U \subset \mathbb{R}^N \rightarrow \mathbb{R}$ et $C \subset U$ convexe. On dit que f est **FORTEMENT CONVEXE** sur C si il existe $\mu > 0$ tel que

$$\forall \alpha \in [0, 1], \forall (x, y) \in C^2, \quad f((1 - \alpha)x + \alpha y) + \frac{\mu}{2}\alpha(1 - \alpha)\|x - y\|^2 \leq (1 - \alpha)f(x) + \alpha f(y).$$

Dans ce cas on dit aussi que f est μ -convexe sur C , et que μ est le coefficient de forte convexité de f sur C . On notera $\Gamma_\mu(C)$ l'ensemble des fonctions fortement convexes sur C .

Remarque III.29. Lorsque $\mu = 0$, on retombe sur la définition de convexité.

Proposition III.30. Soit $f(x) = g(x) + \frac{\mu}{2}\|x\|^2$. Alors $f \in \Gamma_\mu(C)$ si et seulement si $g \in \Gamma_0(C)$. Autrement dit, toute fonction fortement convexe est la somme d'une fonction convexe et d'une norme au carré.

Démonstration. Ici on note comme précédemment $z_\alpha = (1 - \alpha)x + \alpha y$:

$$\begin{aligned} & f \in \Gamma_\mu(C) \\ \Leftrightarrow & \forall \alpha \forall x, y, \quad f(z_\alpha) + \frac{\mu}{2}\alpha(1 - \alpha)\|x - y\|^2 \leq (1 - \alpha)f(x) + \alpha f(y) \\ \Leftrightarrow & \forall \alpha \forall x, y, \quad g(z_\alpha) + \frac{\mu}{2}\|z_\alpha\|^2 + \frac{\mu}{2}\alpha(1 - \alpha)\|x - y\|^2 \\ & \leq (1 - \alpha)g(x) + \alpha g(y) + (1 - \alpha)\frac{\mu}{2}\|x\|^2 + \alpha\frac{\mu}{2}\|y\|^2. \end{aligned}$$

Si on regroupe tous les termes proportionnels à μ , on voit que :

$$\begin{aligned}
 & \frac{1}{2}\|z_\alpha\|^2 + \frac{1}{2}\alpha(1-\alpha)\|x-y\|^2 - (1-\alpha)\frac{1}{2}\|x\|^2 - \alpha\frac{1}{2}\|y\|^2 \\
 = & (1-\alpha)^2\|x\|^2 + \alpha^2\|y\|^2 + 2\alpha(1-\alpha)\langle x, y \rangle + \alpha(1-\alpha)\|x\|^2 + \alpha(1-\alpha)\|y\|^2 - 2\alpha(1-\alpha)\langle x, y \rangle \\
 - & (1-\alpha)\|x\|^2 - \alpha\|y\|^2 \\
 = & \|x\|^2 \left((1-\alpha)^2 + \alpha(1-\alpha) - (1-\alpha) \right) + \|y\|^2 \left(\alpha^2 + \alpha(1-\alpha) - \alpha \right) \\
 = & 0.
 \end{aligned}$$

Donc tous les termes en μ disparaissent, et ce qui reste est exactement la définition pour g d'être convexe. ■

Proposition III.31. *La somme d'une fonction fortement convexe et d'une fonction convexe est fortement convexe.*

Démonstration. Laissé en exercice. ■

Proposition III.32. *La composition d'une fonction fortement convexe avec une application affine injective est fortement convexe.*

Démonstration. Laissé en exercice. ■

III.II.2 Caractérisation de la forte convexité

Proposition III.33 (Forte convexité via Hessienne). *Soit $f : U \subset \mathbb{R}^N \rightarrow \mathbb{R}$, deux fois différentiable sur U , et $C \subset U$ convexe et ouvert. Alors les propriétés suivantes sont équivalentes, pour $\mu > 0$:*

- i) f est fortement convexe sur C , c-à-d $f \in \Gamma_\mu(C)$;
- ii) $(\forall x \in C) \quad \lambda_{\min}(\nabla^2 f(x)) \geq \mu$.

Démonstration. Soit $\mu > 0$ et $f = g + (\mu/2)\|\cdot\|^2$. En particulier on a $\nabla^2 f(x) = \nabla^2 g(x) + \mu I$ sur C . Donc $\lambda_{\min}(\nabla^2 f(x)) = \lambda_{\min}(\nabla^2 g(x)) + \mu$. On conclut donc avec les Propositions III.30 et III.19. ■

Remarque III.34. La forte convexité requiert donc une borne inférieure uniforme sur les valeurs propres de la Hessienne. Au contraire de la stricte convexité qui n'a besoin que de la définie positivité en (« presque ») tout point. Il est essentiel ici de bien faire la distinction entre la caractérisation de la forte convexité :

$$(\exists \mu > 0)(\forall x \in C) \quad \lambda_{\min}(\nabla^2 f(x)) \geq \mu,$$

et la propriété beaucoup plus faible :

$$(\forall x \in C)(\exists \mu > 0) \quad \lambda_{\min}(\nabla^2 f(x)) \geq \mu,$$

qui est en fait équivalente à

$$(\forall x \in C) \quad \lambda_{\min}(\nabla^2 f(x)) > 0,$$

qui implique la stricte convexité seulement.

Exemple III.35. $f(x) = e^x$ est strictement convexe mais n'est pas fortement convexe. On le voit par exemple en notant que f n'est pas coercive, ou bien que f'' tend vers 0 en $-\infty$.

Proposition III.36. Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ une fonction quadratique : $f(x) = \langle Ax, x \rangle + \langle b, x \rangle + c$. Alors f est fortement convexe si et seulement si $A \in \mathcal{M}_N(\mathbb{R})$ est définie positive.

Démonstration. cf. TD



III.II.3 Forte convexité et minimiseurs

Théorème III.37. Toute fonction fortement convexe est coercive.

Démonstration. On va supposer³ par simplicité qu'il existe un point $x_0 \in \mathbb{R}^N$ tel que f soit différentiable en x_0 . D'après la Proposition III.30, on peut écrire $f = g + \frac{\mu}{2} \|\cdot\|^2$, où $g \in \Gamma_0(C)$, et g est différentiable en x_0 par hypothèse. D'après la Proposition III.18, on a également

$$(\forall x \in C) \quad g(x) \geq g(x_0) + \langle \nabla g(x_0), x - x_0 \rangle.$$

On en déduit, via l'inégalité de Cauchy-Schwartz et l'inégalité triangulaire :

$$(\forall x \in C) \quad f(x) \geq g(x_0) - \|\nabla g(x_0)\|(\|x\| + \|x_0\|) + \frac{\mu}{2}\|x\|^2.$$

Comme le membre de droite est un polynôme d'ordre 2 en $\|x\|$, dont le coefficient principal est strictement positif, on en déduit qu'il tend vers $+\infty$ lorsque $\|x\| \rightarrow +\infty$. D'où le résultat.



Corollaire III.38. Soit $f : U \rightarrow \mathbb{R}$ une fonction continue et fortement convexe sur $C \subset U$ fermé. Alors f admet un unique minimiseur global sur C .

Démonstration. D'après le Théorème III.37 f est coercive, donc on peut appliquer le Théorème II.35 et déduire l'existence d'un minimiseur. L'unicité va également découler de la forte

³Le résultat reste vrai sans cette hypothèse ! Mais pour le prouver on aurait besoin d'autres outils. Au choix : Montrer que les fonctions convexes sont localement Lipschitziennes, et donc différentiables presque partout (Théorème de Rademacher) ; Utiliser le Théorème de Hahn-Banach pour séparer l'épigraphhe d'un point quelconque sous l'épigraphhe, et en déduire l'existence d'une minorante affine ; Projeter ce point sur l'épigraphhe et utiliser la caractérisation variationnelle de la projection (cf. dernier chapitre).

convexité. En effet, s'il existait deux minimiseurs x_1^*, x_2^* , on aurait via la Définition III.28 que

$$\frac{1}{2}f(x_1^*) + \frac{1}{2}f(x_2^*) \geq f\left(\frac{x_1^* + x_2^*}{2}\right) + \frac{\mu}{8}\|x_1^* - x_2^*\|^2,$$

où $\frac{1}{2}f(x_1^*) + \frac{1}{2}f(x_2^*) = \min_C f$ par définition de x_1^*, x_2^* , et $f\left(\frac{x_1^* + x_2^*}{2}\right) \geq \min_C f$. Ceci implique donc que $\frac{\mu}{8}\|x_1^* - x_2^*\|^2 \leq 0$, c-à-d que $x_1^* = x_2^*$. ■

III.III Récapitulatif du Chapitre

Ici $C \subset U \subset \mathbb{R}^N$, où U est un ouvert de \mathbb{R}^N , et C est une contrainte fermée non vide. On considère une fonction $f : U \rightarrow \mathbb{R}$, et le problème d'optimisation associé

$$\text{minimiser}_{x \in C} f(x).$$

Unicité des minimiseurs

- Si f est fortement convexe et C convexe alors f admet un unique minimiseur global sur C .

La convexité donne une réciproque au Théorème de Fermat Si f est convexe sur C et C convexe, et que $x \in \text{int } C$, alors ces propriétés sont équivalentes :

- x est un minimiseur global de f sur C
- x est un minimiseur local de f sur C
- $\nabla f(x) = 0$.

Utiliser la Hessienne

- Si C convexe et ouvert, alors f est convexe sur C si et seulement si

$$(\forall x \in C) \quad \lambda_{\min}(\nabla^2 f(x)) \geq 0.$$

- Si C convexe et ouvert, et $\mu > 0$, alors f est μ -fortement convexe sur C si et seulement si

$$(\forall x \in C) \quad \lambda_{\min}(\nabla^2 f(x)) \geq \mu.$$

Chapitre IV

Algorithmes de minimisation sans contrainte

Dans tout ce chapitre, nous allons considérer une fonction différentiable $f : \mathbb{R}^N \rightarrow \mathbb{R}$, que l'on supposera convexe sauf mention du contraire. Rappelons dans ce cas (cf. Théorème III.25) que tout minimiseur $\bar{x} \in \operatorname{argmin} f$ est caractérisé par

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

Cependant, en général, il n'est pas possible de déterminer une formule explicite pour \bar{x} à partir de $\nabla f(\bar{x}) = 0$, car ces équations peuvent être *non linéaires*. C'est pourquoi en pratique on est amené à chercher une *valeur approchée* de \bar{x} . C'est tout l'objet de ce chapitre que de présenter une classe de méthodes classiques pour obtenir de telles solutions approchées : les algorithmes itératifs.

IV.I Méthodes de descente

IV.I.1 Algorithmes itératifs

Comme son nom l'indique, le but d'une méthode itérative est de générer une suite de vecteurs $(x_k)_{k \in \mathbb{N}} \subset \mathbb{R}^N$ telle que, lorsque $k \rightarrow +\infty$, x_k converge vers la solution de notre problème. On peut définir de manière formelle ce qu'est une méthode itérative :

Définition IV.1. Un **ALGORITHME ITÉRATIF** d'ordre $p \geq 1$ sur \mathbb{R}^N est la donnée d'une instruction $\mathbb{A} : (\mathbb{R}^N)^p \rightarrow \mathbb{R}^N$, telle que le nouvel itéré dépende des p itérés précédents

$$(\forall k \in \mathbb{N}) \quad x_{k+1} = \mathbb{A}(x_k, \dots, x_{k-p+1}).$$

On dit alors que $(x_k)_{k \in \mathbb{N}}$ est générée par l'algorithme \mathbb{A} .

En particulier, un algorithme itératif est dit **DU PREMIER ORDRE** sur \mathbb{R}^N si, à chaque itération, le nouvel itéré ne dépend que du précédent; c'est-à-dire qu'il existe une application $\mathbb{A} : \mathbb{R}^N \rightarrow \mathbb{R}^N$ telle que $x_{k+1} = \mathbb{A}(x_k)$.

Exemple IV.2. Les suites arithmétique $x_{k+1} = x_k + r$ ou géométrique $x_{k+1} = rx_k$ sont définies par des algorithmes itératifs du premier ordre sur \mathbb{R} (ici $r \in \mathbb{R}$).

Exemple IV.3. La suite de Fibonacci définie par

$$x_0 = 0, x_1 = 1, \quad x_{k+1} = x_k + x_{k-1}$$

est générée par un algorithme itératif du deuxième ordre sur \mathbb{R} . Par contre elle n'est pas générée par un algorithme itératif du premier ordre sur \mathbb{R} .

Toute méthode du premier ordre peut se réécrire sous la forme

$$x_{k+1} = x_k + \rho_k d_k, \quad \rho_k > 0, \quad d_k \in \mathbb{R}^N, \quad (\text{IV.1})$$

où ρ_k et d_k dépendent de x_k . On dit alors que d_k est la **direction** de l'algorithme au k -ème itéré, et que ρ_k est le **pas** de l'algorithme. Le choix et le rôle donnés à ρ_k et d_k dépendent de l'algorithme.

Remarque IV.4. Faisons le point, et listons ce que l'on peut espérer d'un tel algorithme dans le cadre de notre problème d'optimisation :

- Comme on l'a dit, on souhaite que $\lim_k x_k = x^* \in \operatorname{argmin} f$. C'est la convergence des itérés de la suite vers une solution.
- Au vu de la définition IV.1, et si ρ_k ne tend pas vers 0, on voit que d_k doit tendre vers 0. Or on souhaite à la limite avoir $\nabla f(x) = 0$. Donc il est raisonnable que d_k soit construit à base d'informations sur les dérivées partielles de f .
- On peut également souhaiter la convergence de la suite des valeurs : $\lim_k f(x_k) = \inf f$. De plus, puisque en pratique on va s'arrêter avec k fini, on peut espérer qu'à chaque itération les valeurs s'améliorent, c'est-à-dire $f(x_{k+1}) \leq f(x_k)$.
- On peut également vouloir en savoir plus sur la convergence, d'un point de vue *quantitatif*. Par exemple la VITESSE DE CONVERGENCE des itérés vers une solution, ou des valeurs vers $\inf f$, ou de $\|\nabla f(x_k)\|$ vers 0. On distingue généralement trois « classes » de vitesses :

Définition IV.5. Soit $(r_k)_{k \in \mathbb{N}} \subset [0, +\infty[$ une suite qui tend vers 0 lorsque $k \rightarrow +\infty$. On dit que

- r_k converge **LINÉAIREMENT** si

$$(\exists \theta \in [0, 1[)(\forall k \in \mathbb{N}) \quad r_{k+1} \leq \theta r_k.$$

- r_k converge **SUPERLINÉAIREMENT** si

$$(\exists \theta \in [0, 1[)(\exists \beta \in]1, +\infty[)(\forall k \in \mathbb{N}) \quad r_{k+1} \leq \theta r_k^\beta.$$

- r_k converge **Souslinéairement** si

$$(\exists C \in [0, 1[)(\exists \alpha \in]0, +\infty[)(\forall k \in \mathbb{N}) \quad r_k \leq \frac{C}{k^\alpha}.$$

Remarque IV.6. La convergence linéaire est parfois appelée convergence GÉOMÉTRIQUE, pour des raisons évidentes. Une suite convergeant linéairement vérifie en particulier que

$$r_k \leq \theta^k r_0,$$

c'est-à-dire qu'elle converge exponentiellement.

Remarque IV.7. La convergence superlinéaire est plus rapide que la convergence linéaire. Par récurrence, on voit qu'une telle suite vérifie (rappelons que $r_k \rightarrow 0$)

$$r_k \leq \theta^{\sum_i^k \beta^i} r_0^{\beta^k}.$$

Donc à partir d'un certain rang, la suite tend vers 0 à une vitesse r^{β^k} ce qui est très rapide ! Pour $\beta = 2$ on parle de convergence QUADRATIQUE, c'est en général le mieux que l'on puisse espérer.

Remarque IV.8. La convergence souslinéaire est moins rapide que la convergence linéaire.

IV.I.2 Directions de descente

On s'intéresse ici aux méthodes itératives d'ordre 1 $x_{k+1} = x_k + \rho_k d_k$, et on va s'intéresser à des choix particuliers de d_k qui permettent de garantir que l'algorithme converge vers un minimiseur de la fonction.

Pour cela, on va commencer par répondre à la question : comment s'assurer que

$$f(x_{k+1}) < f(x_k) ?$$

Définition IV.9. Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiable, et $x \in \mathbb{R}^N$. On dit que $d \in \mathbb{R}^N$ est une **DIRECTION DE DESCENTE** en x si la dérivée directionnelle en x dans la direction d est strictement négative :

$$\frac{\partial f}{\partial d}(x) < 0.$$

Remarque IV.10. Rappelons d'après la Proposition I.63 que cela équivaut à $\langle \nabla f(x), d \rangle < 0$, c'est-à-dire former un angle strictement obtus avec $\nabla f(x)$.

Proposition IV.11 (Existence de directions de descente). Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiable, et $x \in \mathbb{R}^N$. Alors il existe une direction de descente en x si et seulement si x n'est pas un point critique.

Démonstration. Si x n'est pas un point critique, i.e. $\nabla f(x) \neq 0$, alors avec $d = -\nabla f(x)$ on a $\langle \nabla f(x), d \rangle = -\|\nabla f(x)\|^2 < 0$. Si x admet une direction de descente d , alors $\nabla f(x)$ ne peut être égal à 0 sinon on aurait $\langle \nabla f(x), d \rangle = 0$. ■

Proposition IV.12 (Décroissance d'Armijo pour les directions de descente). *Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiable, $x \in \mathbb{R}^N$, et d une direction de descente en x . Alors :*

- 1) $(\forall \beta \in]0, 1[)(\exists \rho > 0)(\forall t \in]0, \rho[) \quad f(x + td) \leq f(x) + t\beta \langle \nabla f(x), d \rangle$.
- 2) $(\exists \rho > 0)(\forall t \in]0, \rho[) \quad f(x + td) < f(x)$.

Démonstration. Au vu de la définition de direction de descente, on voit que i) implique trivialement ii). Donc il suffit maintenant de vérifier i). Soit donc $\beta \in]0, 1[$ quelconque. D'après Proposition I.63, on a

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

Donc, d'après la définition de la limite, il existe $\rho > 0$ tel que pour tout $|t| < \rho$,

$$\frac{f(x + td) - f(x)}{t} < \beta \langle \nabla f(x), d \rangle.$$

■

Cette proposition suggère donc que les directions de descente sont des candidates de directions d_k à suivre dans notre algorithme IV.1, puisqu'elle permettent de faire décroître les valeurs de la fonction, pourvu que le pas choisi soit suffisamment petit.

Définition IV.13. Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiable. Une **MÉTHODE DE DESCENTE** pour f est un algorithme itératif du premier ordre de la forme (IV.1), où d_k est une direction de descente en x_k .

La plupart des résultats concernant les directions de descente que l'on vient de voir peuvent s'interpréter de manière géométrique. On peut donc s'aider d'un dessin pour comprendre de quoi il s'agit.

Considérons une fonction $f : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiable, et $x \in \mathbb{R}^N$. On peut alors définir son ENSEMBLE DE NIVEAU en $f(x)$

$$[f = f(x)] := \{x' \in \mathbb{R}^N \mid f(x') = f(x)\}.$$

ENSEMBLE DE SOUS-NIVEAU en $f(x)$ (voir Figure IV.1) :

$$[f \leq f(x)] := \{x' \in \mathbb{R}^N \mid f(x') \leq f(x)\}.$$

On a alors le résultat suivant (énoncé informellement, voir le prochain Chapitre pour plus de détails) :

Théorème IV.14. Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiable, et $x \in \mathbb{R}^N$ un point non critique de f . Alors :

- 1) L'espace tangent à $[f = f(x)]$ est égal à l'ensemble des directions $d \in \mathbb{R}^N$ dont la dérivée directionnelle $Df(x)(d)$ s'annule.
- 2) L'espace normal à $[f = f(x)]$ est la droite vectorielle engendrée par $\nabla f(x)$.

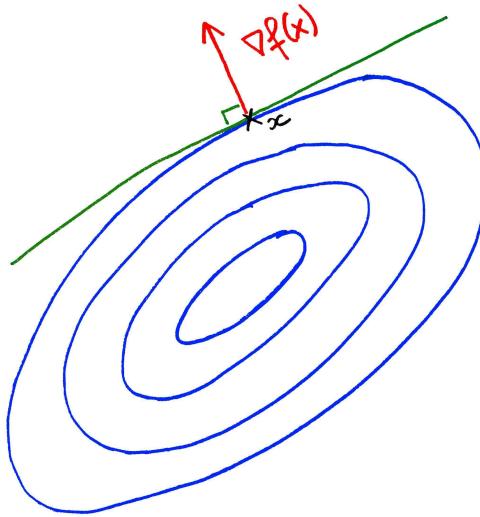


FIGURE IV.1 – Le gradient est normal aux ensembles de sous-niveau et pointe vers l’extérieur.

On peut voir que :

- 1) Le gradient $\nabla f(x)$ est perpendiculaire à la courbe de niveau et « pointe » vers l’extérieur.
- 2) Les directions de descente sont les vecteurs qui « pointent » vers l’intérieur de l’ensemble de sous-niveau.
- 3) Quelque soit la direction de descente suivie, on n’a la descente garantie que si le pas est suffisamment petit. Il faut bien sur déterminer qu’est-ce que ça veut dire en pratique (cf. prochaine section).

IV.I.3 Méthodes du gradient et de Newton

On va ici maintenant à définir des méthodes de descente. On a vu dans la preuve de la Proposition IV.11 que $-\nabla f(x)$ est une direction de descente en x . Ceci nous pousse donc naturellement à définir la méthode du gradient, que l’on étudiera en détail dans la prochaine section :

Définition IV.15. Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiable. La **MÉTHODE DU GRADIENT** est la méthode de descente où l'on choisit $d_k = -\nabla f(x_k)$, c'est-à-dire :

$$x_{k+1} = x_k - \rho_k \nabla f(x_k), \quad \rho_k > 0.$$

On pourrait se demander si cette méthode est bonne, et si l'on peut trouver mieux. Par exemple, on a vu dans la Proposition IV.12.i) que plus la dérivée directionnelle $\langle \nabla f(x), d \rangle$ est négative, et plus on pourra faire décroître les valeurs de la fonction dans cette direction. Il est donc naturel de chercher la direction d qui minimise la dérivée directionnelle en x . On peut en fait montrer que c'est exactement $-\nabla f(x)$, ce qui explique qu'on dise parfois que $-\nabla f(x)$ est la DIRECTION DE LA PLUS GRANDE PENTE :

Proposition IV.16. Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiable, et $x \in \mathbb{R}^N$ un point non critique. Alors

$$\frac{-\nabla f(x)}{\|\nabla f(x)\|} \in \underset{\|d\|=1}{\operatorname{argmin}} \langle \nabla f(x), d \rangle.$$

Démonstration. D'après l'inégalité de Cauchy-Schwarz, on a pour tout $\|d\| = 1$:

$$\langle \nabla f(x), d \rangle \geq -\|\nabla f(x)\| \|d\| = -\|\nabla f(x)\|.$$

Par ailleurs, cette borne inférieure est atteinte si on prend $d = \frac{-\nabla f(x)}{\|\nabla f(x)\|}$. C'est donc par définition un minimiseur de $d \mapsto \langle \nabla f(x), d \rangle$. ■

La Proposition IV.16 nous fournit également une nouvelle interprétation de la méthode du gradient : faire un pas de la méthode du gradient à partir d'un point x , c'est équivalent à minimiser l'approximation de Taylor de f en x à l'ordre 1 sur un voisinage de x . Plus précisément :

Proposition IV.17. Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiable, et $x \in \mathbb{R}^N$ un point non critique. Soit $\rho > 0$, et $x^+ = x - \rho \nabla f(x)$ le point obtenu après avoir fait un pas de la méthode du gradient en partant de x . Alors

$$x^+ \in \underset{x' \in \mathbb{B}(x, \rho \|\nabla f(x)\|)}{\operatorname{argmin}} f(x) + \langle \nabla f(x), x' - x \rangle.$$

Démonstration. D'après l'inégalité de Cauchy-Schwarz, on a pour tout $x' \in \mathbb{B}(x, \rho \|\nabla f(x)\|)$:

$$\langle \nabla f(x), x' - x \rangle \geq -\|\nabla f(x)\| \|x' - x\| \geq -\rho \|\nabla f(x)\|^2.$$

Par ailleurs, cette borne inférieure est atteinte si on prend $x' = x - \rho \nabla f(x)$. C'est donc par définition un minimiseur de $x' \mapsto f(x) + \langle \nabla f(x), x' - x \rangle$. ■

On voit donc que la méthode du gradient exploite au mieux l'information du premier ordre de f en x pour trouver une direction de descente optimale. Du coup il est légitime de se demander ce que l'on obtient lorsque on minimise l'approximation de Taylor de f en x au deuxième ordre. C'est résumé dans le résultat suivant :

Proposition IV.18. Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ deux fois différentiable, et $x \in \mathbb{R}^N$ un point non critique, tel que $\nabla^2 f(x) \succ 0$. Alors

$$x - \nabla^2 f(x)^{-1} \nabla f(x) = \underset{x' \in \mathbb{R}^N}{\operatorname{argmin}} f(x) + \langle \nabla f(x), x' - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(x' - x), (x' - x) \rangle.$$

De plus, $-\nabla^2 f(x)^{-1} \nabla f(x)$ est une direction de descente pour f en x .

Démonstration. On est en train de minimiser la fonction (prendre garde au fait que x est une constante ici !)

$$\phi(x') := f(x) + \langle \nabla f(x), x' - x \rangle + \frac{1}{2} \langle \nabla^2 f(x)(x' - x), (x' - x) \rangle.$$

On voit que c'est une fonction quadratique, telle que

$$\nabla \phi(x') = \nabla f(x) + \nabla^2 f(x)(x' - x) \quad \text{et} \quad \nabla^2 \phi(x') = \nabla^2 f(x).$$

Par hypothèse $\nabla^2 f(x)$ est définie positive donc ϕ est fortement convexe (voir Proposition III.36). Donc elle admet un unique minimiseur (voir Théorème III.38) que l'on notera x^+ . Par convexité de ϕ , ce minimiseur x^+ est caractérisé par la condition d'optimalité du premier ordre $\nabla \phi(x^+) = 0$, qui devient ici

$$\nabla^2 f(x)(x^+ - x) + \nabla f(x) = 0.$$

Puisque on a supposé que $\nabla^2 f(x)$ est inversible, on trouve que la solution de ce système linéaire est $x^+ = x - \nabla^2 f(x)^{-1} \nabla f(x)$. Pour voir que $d = -\nabla^2 f(x)^{-1} \nabla f(x)$ est une direction de descente, on utilise la Proposition I.39 :

$$\begin{aligned} \langle \nabla f(x), -\nabla^2 f(x)^{-1} \nabla f(x) \rangle &= -\langle \nabla^2 f(x) \nabla^2 f(x)^{-1} \nabla f(x), \nabla^2 f(x)^{-1} \nabla f(x) \rangle \\ &\leq -\lambda_{\min}(\nabla^2 f(x)) \|\nabla^2 f(x)^{-1} \nabla f(x)\|^2 < 0. \end{aligned}$$

■

On peut donc définir une nouvelle méthode de descente :

Définition IV.19. Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ deux fois différentiable, et telle que $\nabla^2 f(x) \succ 0$ pour tout $x \in \mathbb{R}^N$. La **MÉTHODE DE NEWTON** est la méthode de descente où l'on choisit $d_k = -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$ et $\rho_k = 1$, c'est-à-dire :

$$x_{k+1} = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k).$$

Remarque IV.20 (Newton). Quelques observations sur la méthode de Newton :

- La méthode de Newton requiert le calcul de la Hessienne de f (ce qui peut coûter cher), et son inversion (ce qui peut coûter encore plus cher).

- Beaucoup de méthodes très efficaces sont définies en remplaçant $\nabla^2 f(x_k)^{-1}$ par une matrice H_k qui est une approximation facile à calculer de $\nabla^2 f(x_k)^{-1}$. Cette famille de méthodes s'appelle les méthodes de Quasi-Newton (voir exercice IV.21).
- On n'étudiera pas cet algorithme, dont l'analyse est compliquée. Plus de détails en M1 dans l'UE Optimisation (OP8). On peut néanmoins citer (cf. TP) que 1) l'algorithme est très sensible aux conditions initiales (choix de x_0) et que 2) quand l'algorithme fonctionne, il converge très vite (plus précisément : superlinéairement).

Exercice IV.21 (Une méthode de Quasi-Newton). Soit $f \in C^2(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N) \cap \Gamma_\mu(\mathbb{R}^N)$. On considère la méthode $x_{k+1} = x_k - D_k^{-1} \nabla f(x_k)$, où $D_k \in \mathcal{M}_N(\mathbb{R})$ est une matrice diagonale dont les coefficients valent $D_{i,i} = \frac{\partial^2 f}{\partial x_i^2}(x_k)$.

- 1) Montrer que $\text{spec}(D_k) \subset [\mu, L]$.
- 2) En déduire que $d_k := -D_k^{-1} \nabla f(x_k)$ est une direction de descente pour f en x_k si l'il n'est pas un point critique.

Pour conclure, il est intéressant de noter que la méthode du gradient, tout comme la méthode de Newton, peut se voir comme la minimisation d'une approximation quadratique de f . Mais ici on parle d'une approximation quadratique qui *ignore l'information du second ordre* de f :

Exercice IV.22 (Une autre caractérisation de la méthode du gradient). Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiable, $\rho > 0$ et $x \in \mathbb{R}^N$ un point non critique. Montrer que

$$x - \rho \nabla f(x) = \underset{x' \in \mathbb{R}^N}{\operatorname{argmin}} f(x) + \langle \nabla f(x), x' - x \rangle + \frac{1}{2\rho} \|x' - x\|^2.$$

IV.II Conditionnement des fonctions convexes à gradient Lipschitzien

IV.II.1 Fonctions à gradient Lipschitzien

Définition IV.23. Soit $F : \mathbb{R}^N \rightarrow \mathbb{R}^M$. On dit que F est **LIPSCHITZIENNE** si

$$(\exists L \in [0, +\infty[)(\forall x, y \in \mathbb{R}^N) \quad \|F(x) - F(y)\| \leq L \|x - y\|.$$

Dans ce cas, on dira parfois que F est L -Lipschitzienne.

Remarque IV.24. On notera $\text{Lip}(F)$ la meilleure (la plus petite) constante de Lipschitz possible pour F . Elle se définit comme :

$$\text{Lip}(F) := \sup_{x \neq y \in \mathbb{R}^N} \frac{\|F(x) - F(y)\|}{\|x - y\|} \in [0, +\infty].$$

On voit alors immédiatement que F est Lipschitzienne si et seulement si $\text{Lip}(F) < +\infty$, ce qui implique en particulier que F est $\text{Lip}(F)$ -Lipschitzienne.

Le quotient $\frac{\|F(x) - F(y)\|}{\|x - y\|}$ qui apparaît dans la remarque ci-dessus n'est pas sans rappeler la définition de la différentielle. Ce n'est pas une simple coïncidence : il se trouve que pour les fonctions différentiables, la constante de Lipschitz se calcule directement à partir de la différentielle (plus précisément, à partir de la jacobienne, qui est la matrice de la différentielle) :

Proposition IV.25 (Lipschitz via la jacobienne). *Soit $F : \mathbb{R}^N \rightarrow \mathbb{R}^M$ une application différentiable sur \mathbb{R}^N . Alors :*

$$\text{Lip}(F) = \sup_{x \in \mathbb{R}^N} \|JF(x)\|.$$

Démonstration. Commençons par définir $L := \sup_{x \in \mathbb{R}^N} \|JF(x)\|$, et montrons que $\text{Lip}(F) = L$ avec deux inégalités.

Si $L = +\infty$, on a forcément $\text{Lip}(F) \leq L$. Si $L < +\infty$, alors on peut utiliser l'inégalité des accroissements finis :

$$\|F(x) - F(y)\| \leq \sup_{z \in \mathbb{R}^N} \|JF(z)\| \|x - y\| = L \|x - y\|.$$

On déduit alors que F est L -Lipschitzienne, ce qui veut dire que $\text{Lip}(F) \leq L$.

Si $\text{Lip}(F) = +\infty$, on a forcément $\text{Lip}(F) \geq L$. Si $\text{Lip}(F) < +\infty$, alors F est $\text{Lip}(F)$ -Lipschitzienne. Si on utilise le fait que (cf. Proposition I.63)

$$DF(x)(d) = \lim_{t \rightarrow 0} \frac{F(x + td) - F(x)}{t},$$

on peut écrire pour tout $x \in \mathbb{R}^N$:

$$\|JF(x)\| = \|DF(x)\| = \sup_{\|d\|=1} \|DF(x)(d)\| = \sup_{\|d\|=1} \lim_{t \rightarrow 0} \frac{\|F(x + td) - F(x)\|}{t} \leq \text{Lip}(F).$$

On en déduit que $\text{Lip}(F) \geq L$. ■

Cette proposition nous permet donc de calculer/estimer la constante de Lipschitz d'une application F en pratique. En effet, il suffit de calculer la matrice Jacobienne de F en tout point x , de calculer la norme subordonnée euclidienne de la matrice $JF(x)$, puis de trouver une borne supérieure **uniforme** pour cette norme, au sens où elle soit indépendante de x .

Définition IV.26. On note $C_L^{1,1}(\mathbb{R}^N)$ l'ensemble des fonctions $f : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiables et dont le gradient est L -Lipschitzien.

Proposition IV.27. Soit $f \in \Gamma_0(\mathbb{R}^N) \cap C^2(\mathbb{R}^N)$, et $L > 0$. Alors les propriétés suivantes sont équivalentes :

- i) $f \in C_L^{1,1}(\mathbb{R}^N)$ (autrement dit, ∇f est L -Lipschitzien).
- ii) $(\forall x \in \mathbb{R}^N) \quad \lambda_{\max}(\nabla^2 f(x)) \leq L$.

Démonstration. Soit $F = \nabla f$, qui, par hypothèse, est de classe $C^1(\mathbb{R}^N)$. La Proposition I.78.iii) nous dit que $JF = \nabla^2 f$, et la Proposition I.78.i) nous garantit que la Hessienne est symétrique, ce qui nous permet d'écrire en vertu de la Proposition I.36 pour tout $x \in \mathbb{R}^N$ que $\|\nabla^2 f(x)\| = \rho(\nabla^2 f(x))$. De plus, f est supposée convexe, donc le Théorème III.19 nous garantit que les valeurs propres de la Hessienne sont positives, ce qui veut dire que $\rho(\nabla^2 f(x)) = \lambda_{\max}(\nabla^2 f(x))$. On conclut alors avec la Proposition IV.25. ■

On voit donc ici une propriété en quelque sorte **duale**¹ du Théorème III.33 : une borne uniforme inférieure sur le spectre de la Hessienne équivaut à la forte convexité, tandis qu'ici on voit qu'une forte uniforme supérieure équivaut à la Lipschitzianité du gradient. On en déduit d'ailleurs immédiatement que :

Proposition IV.28. Si $f \in C^2(\mathbb{R}^N) \cap \Gamma_\mu(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ avec $\mu, L > 0$, alors $L \geq \mu$.

Démonstration. C'est une directe conséquence des Propositions IV.27 et III.33. ■

Exercice IV.29 (Constante de Lipschitz). Dans cet exercice nous allons calculer (ou estimer) la constante de lipschitz de $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$, pour certaines fonctions $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Soient $A \in \mathbb{R}^{m \times n}$, et $b \in \mathbb{R}^m$.

- 1) Soit $f(x) = \|Ax - b\|^2$. Calculer la Hessienne de f , puis en déduire la constante de Lipschitz de ∇f .
- 2) Soit $f(x) = g(x) + \frac{\mu}{2}\|x\|^2$, où $g : \mathbb{R}^n \rightarrow \mathbb{R}$ est une fonction de classe C^2 et de gradient L -Lipschitzien. Calculer la constante de Lipschitz de ∇f .
- 3) Soit $f(x) = \frac{1}{m} \sum_{i=1}^m \ln(1 + e^{-b_i \langle x, a_i \rangle})$, où a_i est le vecteur apparaissant à la i -ième ligne de la matrice A , et on suppose ici que $|b_i| = 1$.
 - a) Soit $f_i(x) = \ln(1 + e^{-b_i \langle x, a_i \rangle})$. Calculer son gradient et sa Hessienne.
 - b) Vérifier que pour tout $t \in \mathbb{R}$, $\frac{t}{(1+t)^2} \leq \frac{1}{4}$. En déduire que ∇f_i est L_i -Lipschitzien, avec $L_i \leq \|a_i\|^2/4$.
 - c) En déduire que ∇f est L -Lipschitzien, avec $L \leq \frac{1}{4m} \sum_{i=1}^m \|a_i\|^2$.

¹Il existe d'ailleurs une très jolie théorie de la dualité en analyse convexe qui permet entre autres choses de formellement justifier que « forte convexité » et « différentiable à gradient Lipschitzien » sont les deux facettes d'une même pièce. C'est en quelque sorte un résultat analogue à la correspondance entre « régularité » et « décroissance » via la transformée de Fourier. Mais cela est évidemment hors-programme ...

IV.II.2 Conditionnement d'une fonction

Définition IV.30. Soit $f \in \Gamma_\mu(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$, pour $L \geq \mu > 0$. Le **CONDITIONNEMENT** de f , noté $\text{cond}(f)$, est défini par le ratio $\frac{L}{\mu} \in [1, +\infty[$.

Remarque IV.31. Le fait que le conditionnement soit un nombre plus grand que 1 vient de la Proposition IV.28 qui garantit que $L \geq \mu$.

Exemple IV.32. Soit A une matrice symétrique définie positive, et $f(x) = \frac{1}{2}\langle Ax, x \rangle + \langle b, x \rangle + c$ une fonction quadratique. Alors

$$\text{cond}(f) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} = \text{cond}(A).$$

On retrouve ici la notion de conditionnement d'une matrice $\text{cond}(A)$, qui est très importante en Calcul Matriciel : on sait qu'elle contrôle plusieurs choses comme :

- La **stabilité** des algorithmes par rapport aux erreurs
- La **vitesse de convergence** des méthodes de résolution des systèmes linéaires associés

On verra qu'il se passe la même chose pour les fonctions fortement convexes à gradient Lipschitzien : plus le conditionnement sera proche de 1, et meilleurs seront les résultats.

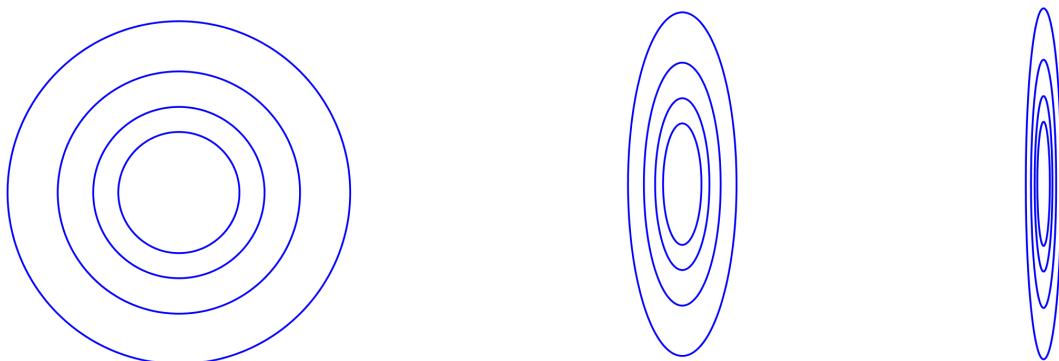


FIGURE IV.2 – Ensembles de niveau pour une fonction quadratique ayant un conditionnement $\text{cond}(f) = 1, 10, 100$ (de gauche à droite).

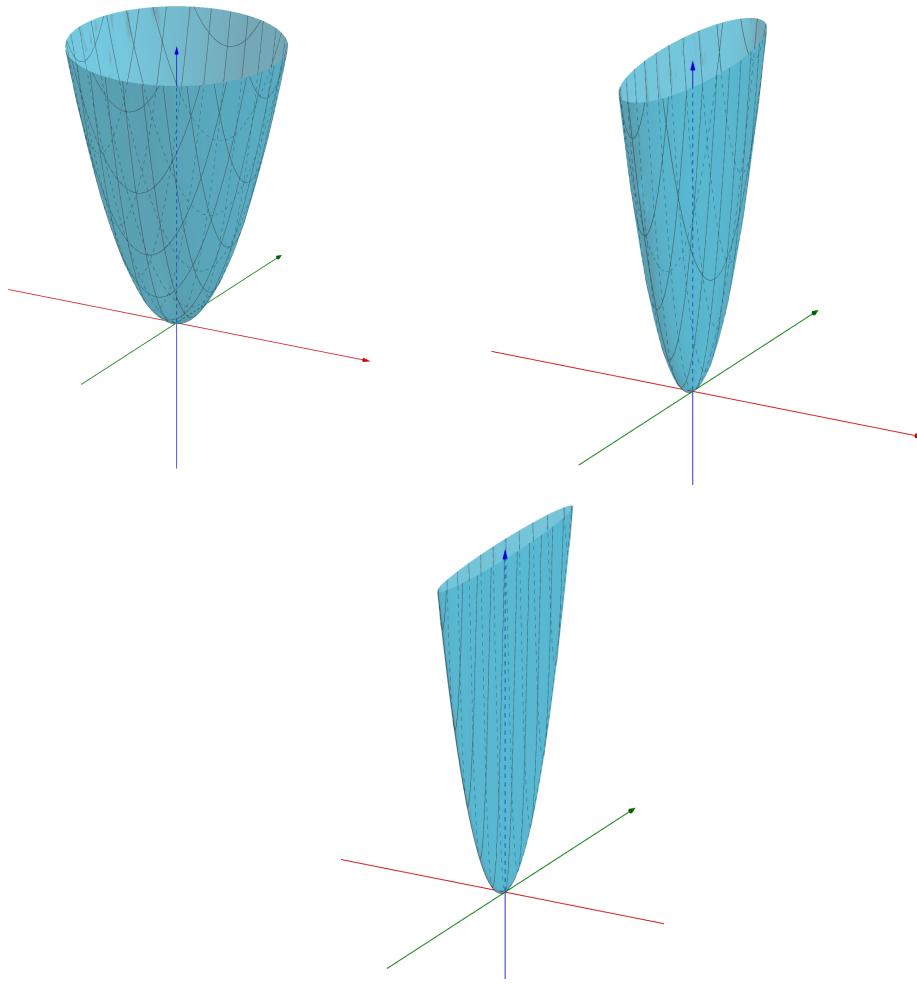


FIGURE IV.3 – Graphe d'une fonction quadratique ayant un conditionnement $\text{cond}(f) = 1, 10, 100$ (de gauche à droite).

Exercice IV.33 (Conditionnement d'une fonction vs. de la Hessienne). Soit $f \in \Gamma_\mu(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N) \cap C^2(\mathbb{R}^N)$, pour $L \geq \mu > 0$. Montrer que

$$\text{cond}(f) \geq \sup_{x \in \mathbb{R}^N} \text{cond}(\nabla^2 f(x)).$$

Trouver un exemple de fonction pour laquelle cette inégalité est stricte.

IV.III Méthode du gradient

IV.III.1 La méthode du gradient à pas fixe

On considère ici l'algorithme du gradient où le pas est fixé tout au long de l'algorithme, c'est à dire

$$x_{k+1} = x_k - \rho \nabla f(x_k), \quad \rho > 0.$$

Dans toute la suite de ce chapitre, on utilisera la notation suivante

$$(\forall x \in \mathbb{R}^N) \quad x^+ := x - \rho \nabla f(x),$$

où x^+ désigne le point que l'on obtient en appliquant un pas de la méthode du gradient à x . Observer que la notation est ambiguë par rapport à la valeur de ρ mais on fera attention à toujours l'utiliser dans un contexte où on sait ce que vaut ρ .

Une question essentielle à propos de cet algorithme est : comment choisir ρ ? On a vu dans la Proposition IV.12 qu'il fallait que ρ soit suffisamment petit pour garantir que l'algorithme fait décroître les valeurs de f . Mais d'un autre côté on imagine bien que si le pas est trop petit, on va faire des tout petits pas, donc l'algorithme va être lent et peu efficace. Il faut donc bien analyser ce qui se passe pour pouvoir prendre le meilleur pas possible.

Proposition IV.34 (Décroissance de la méthode du gradient). Soient $L > 0$, $f \in C_L^{1,1}(\mathbb{R}^N)$ et $\rho > 0$. Soit $x \in \mathbb{R}^N$, et notons $x^+ := x - \rho \nabla f(x)$. Alors :

i) $f(x^+) - f(x) \leq -\rho \left(1 - \frac{L\rho}{2}\right) \|\nabla f(x)\|^2.$

ii) Si $\rho < 2/L$ et x n'est pas un point critique, alors $f(x^+) < f(x)$.

Remarque IV.35 (Choix du pas fixe et conditionnement). La condition $\rho < 2/L$ nous garantit que le pas est *suffisamment petit* pour que la fonction décroisse après un pas de l'algorithme. Mais il faut garder en tête que cette contrainte correspond en quelque sorte à un « pire des cas » : si on prend un pas plus grand, il se peut que quelque part, il y ait un point où l'on va aller « trop loin » et faire réaugmenter les valeurs de la fonctions. En conséquence, cela veut dire aussi que cette condition peut parfois être trop stricte, car il y a des points où on pourrait prendre un pas plus grand. On le voit très bien sur la Figure IV.4, où pour une fonction avec $\text{cond}(f) = 10$, on voit qu'en le point y , le pas $\rho < 2/L$ ne nous permet pas d'aller très loin. Mais on ne peut pas non plus prendre un pas plus grand, car en le point x un pas supérieur à $2/L$ nous ferait sortir du sous-niveau.

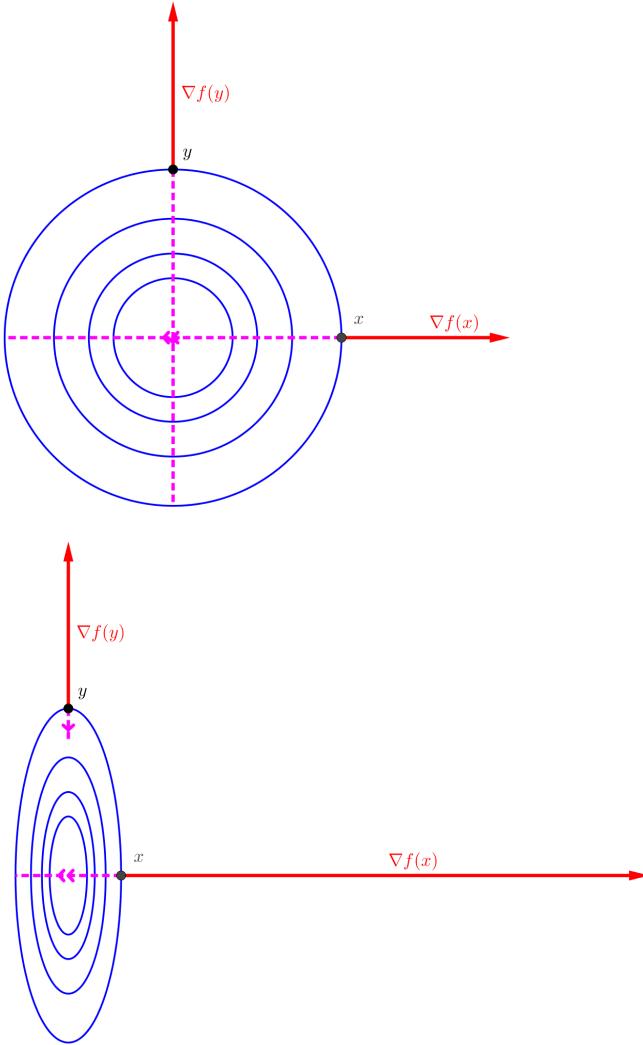


FIGURE IV.4 – Fonctions quadratiques avec un conditionnement de $\text{cond}(f) = 1, 10$ respectivement. En noir, un point x (resp. y) appartenant à l'espace propre de la plus grande (resp. plus petite) valeur propre de la Hessienne. En rouge, les gradients en ces points. En rose, l'ensemble des points que l'on peut atteindre en prenant un pas $\rho < 2/L$.

Remarque IV.36. Pour contourner ce problème mentionné dans la précédente remarque, on pourrait penser à prendre un pas ρ_k qui dépend du point x_k et s'adapte à la géométrie locale de la fonction. On en reparlera dans la prochaine section.

Démonstration de la Proposition IV.34. Soient $x, y \in \mathbb{R}^N$ quelconques. Posons $g(t) = f(z_t)$ où $z_t = (1-t)x + ty$, telle que $g'(t) = \langle \nabla f(z_t), y - x \rangle$. On peut alors écrire :

$$f(y) - f(x) = g(1) - g(0) = \int_0^1 g'(t) dt = \int_0^1 \langle \nabla f(z_t), y - x \rangle dt.$$

Afin de pouvoir utiliser la Lipschitzianité de ∇f , on va faire apparaître un $\nabla f(x)$ puis utiliser l'inégalité de Cauchy-Schwarz :

$$\begin{aligned} f(y) - f(x) &= \int_0^1 \langle \nabla f(z_t) - \nabla f(x), y - x \rangle dt + \langle \nabla f(x), y - x \rangle \\ &\leq \int_0^1 \|\nabla f(z_t) - \nabla f(x)\| \|y - x\| dt + \langle \nabla f(x), y - x \rangle \\ &\leq \int_0^1 L \|z_t - x\| \|y - x\| dt + \langle \nabla f(x), y - x \rangle. \end{aligned}$$

Si on utilise le fait que, par définition, $z_t - x = t(y - x)$, on obtient alors :

$$(\forall x, y \in \mathbb{R}^N) \quad f(y) - f(x) \leq \frac{L}{2} \|y - x\|^2 + \langle \nabla f(x), y - x \rangle. \quad (\text{IV.2})$$

Prenons maintenant $y = x^+ = x - \rho \nabla f(x)$:

$$f(x^+) - f(x) \leq \left(\frac{L}{2} \rho^2 - \rho \right) \|\nabla f(x)\|^2. \quad (\text{IV.3})$$

On conclut en observant que $x \notin \text{crit } f$ garantit $\|\nabla f(x)\|^2 > 0$, et $0 < \rho < \frac{2}{L}$ implique que $(\frac{L}{2} \rho^2 - \rho) < 0$. ■

On a donc vu qu'un pas $\rho \in]0, 2/L[$ est nécessaire pour garantir la décroissance de la fonction le long des itérés de l'algorithme. Mais ceci ne garantit pas la convergence de l'algorithme. Pour cela, on va faire l'hypothèse supplémentaire que la fonction est fortement convexe.

Théorème IV.37 (Convergence linéaire des itérés (cas fortement convexe)). Soient $L \geq \mu > 0$ et $f \in C^2(\mathbb{R}^N) \cap \Gamma_\mu(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$. On note $x^* = \arg\min f$, et on considère la méthode du gradient avec un pas constant $\rho \in]0, 2/L[$. Alors

- i) La suite x_k converge vers x^* .
- ii) La suite $(\|x_k - x^*\|)_{k \in \mathbb{N}}$ converge linéairement, c'est-à-dire que :

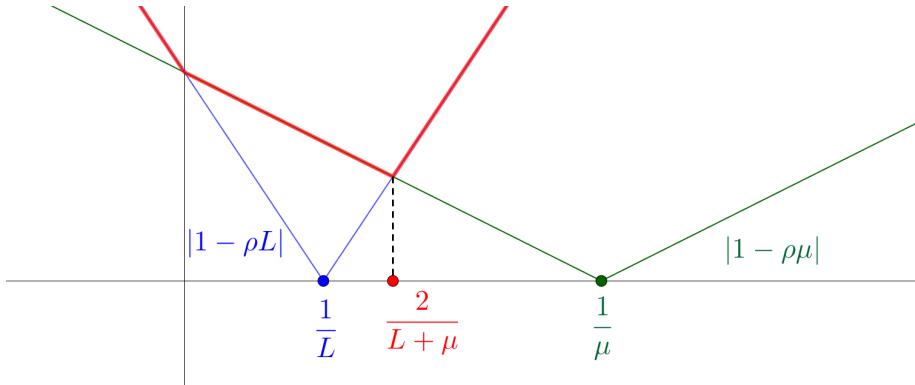
$$(\exists \theta \in [0, 1[)(\forall k \in \mathbb{N}) \quad \|x_{k+1} - x^*\| \leq \theta \|x_k - x^*\|.$$

- iii) Plus précisément, on peut montrer que

$$\theta = \max\{|1 - \rho\mu|; |1 - \rho L|\} = \begin{cases} 1 - \rho\mu & \text{si } \rho \leq \frac{2}{\mu+L} \\ \rho L - 1 & \text{si } \rho \geq \frac{2}{\mu+L}, \end{cases} \quad (\text{IV.4})$$

- iv) Le taux de convergence linéaire θ est minimal lorsque $\rho = 2/(\mu + L)$.

On voit donc que la vitesse de convergence ne dépend que du choix de ρ et du conditionnement de f .



Remarque IV.38 (Pas optimal). La meilleure vitesse est atteinte lorsque θ est le plus petit possible. Au vu de la définition de θ , il est minimal lorsque $\rho = 2/(\mu + L)$, auquel cas $\theta = \frac{L-\mu}{L+\mu}$ (voir aussi Figure IV.III.1). On dit parfois que ce choix de pas est le PAS OPTIMAL. Attention à ne pas confondre avec la Section IV.III.2 ! Il est également possible de montrer que cette vitesse linéaire en $\frac{L-\mu}{L+\mu}$ est la meilleure que l'on puisse espérer avec la méthode du gradient (hors programme). L'inconvénient néanmoins de ce choix de pas est qu'il nécessite la connaissance de μ , ce qui n'est pas toujours le cas en pratique, où L est beaucoup plus facile à estimer.

Exemple IV.39. Il est possible de montrer que pour la fonction quadratique $f(x_1, x_2) = (\mu/2)x_1^2 + (L/2)x_2^2$ et $x = (\sqrt{L}, \sqrt{\mu})$ et $\rho = 2/(\mu + L)$,

$$\|x^+ - x^*\| = \frac{L - \mu}{L + \mu} \|x - x^*\|,$$

donc on ne peut pas améliorer cette vitesse.

Remarque IV.40 (Pas court). Le choix le plus populaire, lorsqu'on ne connaît pas μ , est de prendre $\rho = 1/L$. Dans ce cas, $\theta = 1 - \mu/L$. C'est un choix raisonnable, au sens où il donne la meilleure contraction qu'on puisse garantir avec cet algorithme, lorsqu'on ne connaît pas μ . En effet, sur $]0, 1/L]$, θ est décroissant, tandis que $2/(\mu + L)$ est toujours supérieur à $1/L$, mais peut être arbitrairement proche voire égal à $1/L$. On parle parfois de PAS COURT pour désigner ce choix de pas.

Démonstration du Théorème IV.37. Ici on suppose pour simplifier la preuve que f est également de classe $C^2(\mathbb{R}^N)$, bien que ce ne soit pas nécessaire. Une preuve sans cette hypothèse est disponible dans la Section A.II.1 en Annexe. On cherche donc à montrer que

$$(\exists \theta \in [0, 1[)(\forall x \in \mathbb{R}^N) \quad \|x^+ - x^*\| \leq \theta \|x - x^*\|.$$

On définit le champ de vecteurs associé à l'algorithme : $\mathbb{A} : \mathbb{R}^N \rightarrow \mathbb{R}^N$, $\mathbb{A}(x) = x^+ = x - \rho \nabla f(x)$. En observant que x^* est un point fixe de \mathbb{A} ($\mathbb{A}(x^*) = x^*$), on peut réécrire le problème comme

$$(\exists \theta \in [0, 1])(\forall x \in \mathbb{R}^N) \quad \|\mathbb{A}(x) - \mathbb{A}(x^*)\| \leq \theta \|x - x^*\|.$$

On voit alors qu'il suffit de montrer que \mathbb{A} est Lipschitzienne, avec une constante $Lip(\mathbb{A})$ strictement plus petite que 1. Or f étant C^2 , a fortiori \mathbb{A} est de classe C^1 , et donc on peut utiliser la caractérisation de la Proposition IV.25 qui nous dit que

$$Lip(\mathbb{A}) = \sup_{x \in \mathbb{R}^N} \|J\mathbb{A}(x)\|.$$

Pour tout $x \in \mathbb{R}^N$, on peut calculer $J\mathbb{A}(x) = I - \rho \nabla^2 f(x)$, qui est une matrice symétrique, donc sa norme peut être calculée via ses valeurs propres :

$$Lip(\mathbb{A}) = \sup_{x \in \mathbb{R}^N} \max | \text{spec} \left(I - \rho \nabla^2 f(x) \right) | = \sup_{x \in \mathbb{R}^N} \max_{\lambda \in \text{spec}(\nabla^2 f(x))} |1 - \rho \lambda|.$$

Or on sait via Proposition IV.27 et III.33 que $\text{spec}(\nabla^2 f(x)) \subset [\mu, L]$. Donc nécessairement :

$$(\forall \lambda \in \text{spec}(\nabla^2 f(x))) \quad |1 - \rho \lambda| \leq \max\{|1 - \rho \mu|, |1 - \rho L|\},$$

et on déduit de tout ce qui précède que l'énoncé du Théorème est vrai avec $\theta := \max\{|1 - \rho \mu|, |1 - \rho L|\}$. Il reste maintenant à étudier θ .

Tout d'abord, c'est un simple exercice (non trivial, faire un dessin aide beaucoup, cf. Figure IV.III.1) que de vérifier que

$$\max\{|1 - \rho \mu|, |1 - \rho L|\} = \begin{cases} |1 - \rho \mu| & \text{si } \rho \leq \frac{2}{\mu + L} \\ |\rho L - 1| & \text{si } \rho \geq \frac{2}{\mu + L}. \end{cases}$$

D'autre part, puisque $2/(\mu + L) \in [1/L, 2/L[$, on en déduit que $1 - \rho \mu \in [0, 1[$ et $\rho L - 1 \in [0, 1[$. ■

Remarque IV.41. Ce Théorème IV.37 et sa preuve nécessitent l'hypothèse que f soit de classe C^2 , ce qui nous permet d'exploiter les propriétés de la Hessienne, et des matrices symétriques. Sachez qu'on peut tout à fait se passer de cette double différentiabilité, et simplement supposer que f est convexe, différentiable, et à gradient Lipschitzien. Cela requiert évidemment une preuve différente, qui est un peu plus longue, et que l'on omettra donc ici.

Théorème IV.42 (Convergence linéaire des valeurs (cas fortement convexe)). Soient $L \geq \mu > 0$ et $f \in C^2(\mathbb{R}^N) \cap \Gamma_\mu(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$. On considère la méthode du gradient avec un pas constant $\rho \in]0, 2/L[$. Alors $(f(x_k) - \inf f)_{k \in \mathbb{N}}$ converge linéairement, c'est-à-dire :

$$(\exists \theta \in [0, 1])(\forall k \in \mathbb{N}) \quad f(x_{k+1}) - \inf f \leq \theta^2(f(x_k) - \inf f).$$

Plus précisément, on peut montrer que θ est le même que celui défini dans le Théorème IV.37.iii).

Démonstration. Admis. Une démonstration est disponible dans l'Annexe (Section A.II.1). ■

Exemple IV.43. Soit $f(x) = x^2/2$, tel que $\mu = L = 1$. Alors, pour tout $x \in \mathbb{R}$ et $\rho \in]0, 2[$, on a :

$$f(x^+) = \frac{1}{2}(1 - \rho)^2 x^2 \quad \text{et} \quad f(x) = \frac{1}{2}x^2.$$

Donc on a ici $\theta = (1 - \rho)$. On voit donc que le θ du Théorème est difficilement améliorable.

Pour conclure sur la convergence de la méthode du gradient, il est bon de savoir que même lorsque la fonction n'est pas fortement convexe, l'algorithme du gradient converge. Par contre sa performance est moindre, on passe d'une convergence linéaire pour les valeurs à une convergence souslinéaire. Nous admettons ici sa preuve, qui est un peu longue, étant donné que nous avons déjà bien traité le cas fortement convexe.

Théorème IV.44 (Convergence de la méthode du gradient, cas convexe). Soit $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$, pour $L > 0$, telle que $\operatorname{argmin} f \neq \emptyset$. On considère la méthode du gradient avec un pas constant $\rho \in]0, 2/L[$. Alors

- i) x_k converge vers $x^* \in \operatorname{argmin} f$.
- ii) $f(x_k)$ converge vers $\inf f$.
- iii) Plus précisément, $f(x_k) - \inf f = O\left(\frac{1}{k}\right)$.

Démonstration. Admis. La démonstration complète est disponible dans l'Annexe (voir Section A.II.2). ■

Remarque IV.45. Il n'y a pas de vitesses pour les itérés dans ce Théorème, car ils peuvent tendre vers 0 de manière arbitrairement lente. Pour le voir il suffit de considérer des fonctions qui ressemblent à $f(x) = |x|^p$ pour $p \rightarrow +\infty$.

Remarque IV.46. L'hypothèse $\operatorname{argmin} f \neq \emptyset$ est importante. Si il n'y a pas de minimiseurs, l'algorithme diverge et $f(x_k)$ tend vers $\inf f$ avec une vitesse qui peut être arbitrairement faible. Pour le voir il suffit de considérer des fonctions qui ressemblent à $f(x) = 1/|x|^p$, pour $p \rightarrow +\infty$: dans ce cas la fonction est de plus en plus plate au voisinage de 0, donc le gradient devient très petit, et l'algorithme met de plus en plus de temps à progresser.

Remarque IV.47 (Adaptivité à la forte convexité). On dit que la méthode du gradient à pas constant est **adaptive** à la forte convexité. En effet : si on dispose d'une fonction $f \in C_L^{1,1}(\mathbb{R}^N) \cap \Gamma_0(\mathbb{R}^N)$, alors on peut choisir $\rho < \frac{2}{L}$ et être garantit que l'algorithme va converger, avec une vitesse qui sera au pire de l'ordre de $\frac{1}{t}$ pour les valeurs (Théorème IV.44). Mais ! Si il s'avère que la fonction f est fortement convexe (sans qu'on le sache), alors cet algorithme va converger plus vite que prévu, c'est-à-dire linéairement (Théorème

IV.42). Il est remarquable que l'algorithme soit capable d'exploiter cette propriété de forte convexité sans qu'on ait besoin de le lui dire. C'est pour cela qu'on parle d'adaptivité.

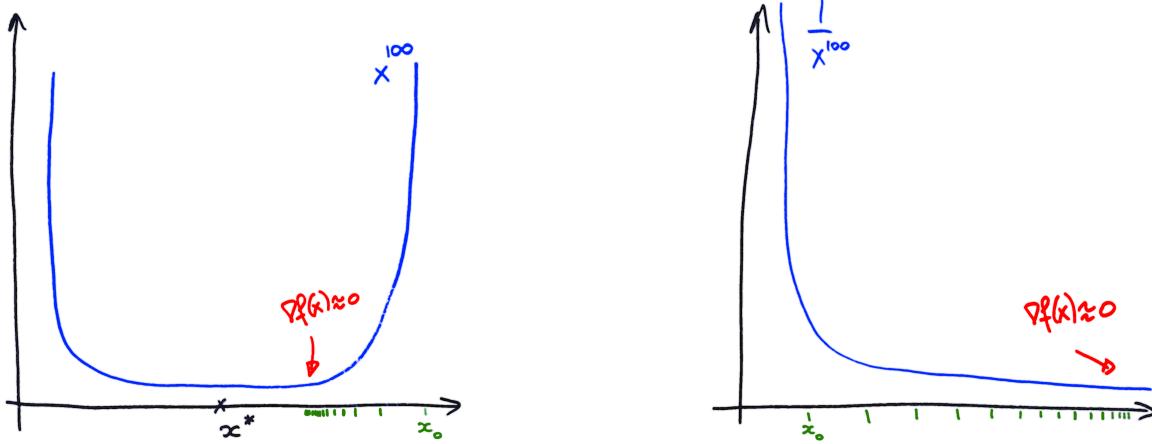


FIGURE IV.5 – Convergence lente de la méthode du gradient pour des fonctions qui s'aplatissent.

IV.III.2 Méthode du gradient à pas optimal

Dans la section précédente, on a vu qu'on pouvait garantir

$$(\forall x \in \mathbb{R}^N) \quad f(x_+) - \inf f \leq \theta^2(f(x) - \inf f),$$

pourvu qu'on choisisse bien ρ . Mais il y a quelques problèmes à cela :

- 1) Pour que cela marche un tant soit peu (c'est-à-dire pour que les valeurs décroissent), la Proposition IV.34 nous dit qu'il faut prendre $\rho < 2/L$. Ce qui nécessite de connaître L , ce qui n'est pas toujours le cas. Idéalement on voudrait une méthode qui ne requière aucune connaissance préalable sur la fonction f : c'est-à-dire qu'elle soit adaptive à L .
- 2) Pour que cela marche bien, il faut prendre le pas optimal $\rho = 2/(\mu + L)$, mais ici encore, μ n'est pas toujours accessible.
- 3) Même si on avait accès à μ et L , nos résultats de contraction des vitesses est vrai en **tout** $x \in \mathbb{R}^N$. Ce qui veut dire que la contraction que l'on a est un « pire des cas », au sens où il y a des mauvais x pour lesquels on va avoir une contraction θ , mais rien n'empêche que pour un autre « bon » x la contraction soit meilleure.

Cela suggère donc que l'on choisisse ρ_k en *boucle ouverte*, c'est-à-dire que le choix de ρ_k va être spécifique à x_k . Une façon de faire est de carrément choisir parmi tous les pas possibles celui qui va donner un point x_{k+1} qui va le plus faire décroître la fonction :

Définition IV.48. L'algorithme du gradient **À PAS OPTIMAL** est défini par

$$x_{k+1} = x_k - \rho_k \nabla f(x_k) \quad \text{où } \rho_k = \operatorname{argmin}_{\rho > 0} f(x_k - \rho \nabla f(x_k)).$$

Remarque IV.49. Ne pas confondre cette méthode du gradient à pas optimal avec la méthode du gradient à pas constant optimal, vu dans la précédente section, où $\rho = 2/(\mu + L)$.

Remarque IV.50. C'est ce que l'on appelle une méthode de recherche en ligne : on cherche le long de l'espace unidimensionnel $\{x - \rho \nabla f(x) \mid \rho \in \mathbb{R}\}$ un bon successeur à x . Il existe de nombreuses autres méthodes de ce type (voir l'exercice suivant).

Exercice IV.51 (Recherche en ligne naïve). On considère la méthode du gradient $x_{k+1} = x_k - \rho_k \nabla f(x_k)$ où $\rho_k > 0$ est calculé à chaque itération selon la règle naïve suivante : on accepte n'importe quelle valeur de ρ_k , pourvu que l'on ait $f(x_{k+1}) < f(x_k)$.

- 1) Prouver la formule suivante pour tout $k \geq 1$: $\prod_{t=2}^{k+1} \left(1 - \frac{1}{t^2}\right) = \frac{1}{2} \frac{k+2}{k+1}$.
- 2) Soit $f(x) = \frac{1}{2}x^2$, et $x_0 \neq 0$. On considère le choix de pas $\rho_k = \frac{1}{k^2}$
 - a) Exprimer x_{k+1} en fonction de x_k . Vérifier que ρ_k respecte notre règle naïve.
 - b) Montrer que x_k converge vers $\frac{x_0}{2}$. Que pouvez-vous en déduire ?
- 3) Même question avec cette fois-ci $\rho_k = 2 - \frac{1}{k^2}$.

Une des propriétés importantes de la méthode du gradient à pas optimal est qu'elle génère des trajectoires en zig-zag :

Proposition IV.52 (Propriété du zig-zag). Soit $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$, pour $L > 0$, telle que $\operatorname{argmin} f \neq \emptyset$. On considère la méthode du gradient à pas optimal. Alors,

$$(\forall k \in \mathbb{N}) \quad \langle \nabla f(x_k), \nabla f(x_{k+1}) \rangle = 0.$$

Démonstration. Soit $x \notin \operatorname{argmin} f$, et $g : [0, +\infty[\rightarrow \mathbb{R}$, $t \mapsto f(x - t \nabla f(x))$. Puisque x n'est pas un minimiseur de f , on a forcément $\nabla f(x) \neq 0$. Donc $-\nabla f(x)$ est une direction de descente, donc d'après la Proposition IV.12, 0 n'est pas un minimiseur de g . Donc ρ est dans l'ouvert $]0, +\infty[$, donc $g'(\rho) = 0$. Or, on peut calculer

$$g'(t) = \langle \nabla f(x - t \nabla f(x)), -\nabla f(x) \rangle. \quad \blacksquare$$

Remarque IV.53. Calculer le pas optimal nécessite donc de résoudre un problème d'optimisation à chaque itération. Pour que ce soit rentable, il faudrait vraiment que l'algorithme soit très efficace, i.e. qu'il converge très rapidement. C'est donc pour cela qu'on va analyser sa convergence plus bas. De toute façon, en pratique :

- On ne minimise pas exactement $f(x_k - \rho \nabla f(x_k))$, mais on cherche un ρ qui soit « pas trop mal », et il y a plein de façons de définir ce que « pas trop mal » veut dire.
- Dans le cas particulier des fonctions quadratiques, on dispose d'une formule explicite pour exprimer ρ_k :

Proposition IV.54. Soit $A \in \mathcal{M}_{M,N}(\mathbb{R})$ inversible, $y \in \mathbb{R}^M$ et $f(x) := \frac{1}{2} \|Ax - y\|^2$. Alors, pour tout $x_k \notin \operatorname{argmin} f$, le pas optimal vaut :

$$\rho_k = \frac{\|\nabla f(x_k)\|^2}{\|A \nabla f(x_k)\|^2}.$$

Remarque IV.55. Si on préfère écrire la fonction quadratique sous la forme $f(x) = \frac{1}{2} \langle Sx, x \rangle + \langle b, x \rangle + c$ avec $S \in \mathcal{S}_N(\mathbb{R})$, alors le pas optimal vaut

$$\rho_k = \frac{\|\nabla f(x_k)\|^2}{\langle S \nabla f(x_k), \nabla f(x_k) \rangle}.$$

Démonstration. On cherche donc à trouver t qui minimise g . Tout d'abord, observons que $g(t) = f(x - t \nabla f(t))$ est la composition d'une fonction fortement convexe avec une fonction affine injective, donc g est fortement convexe. Donc elle admet un unique minimiseur, qu'on note ρ . Puisque f est une fonction quadratique, la propriété du zig-zag $g'(\rho) = 0$ est équivalente à :

$$\begin{aligned} 0 &= \langle A^\top (A(x - \rho A^\top (Ax - b)) - b), A^\top (Ax - b) \rangle \\ &= \|A^\top (Ax - b)\|^2 - \rho \|AA^\top (Ax - b)\|^2, \end{aligned}$$

et la conclusion suit. ■

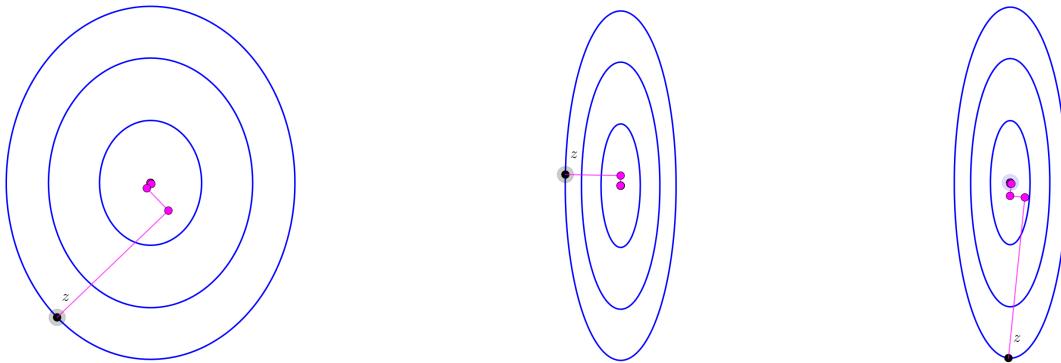


FIGURE IV.6 – Méthode du gradient optimal (GPO) pour diverses fonctions et points initiaux.

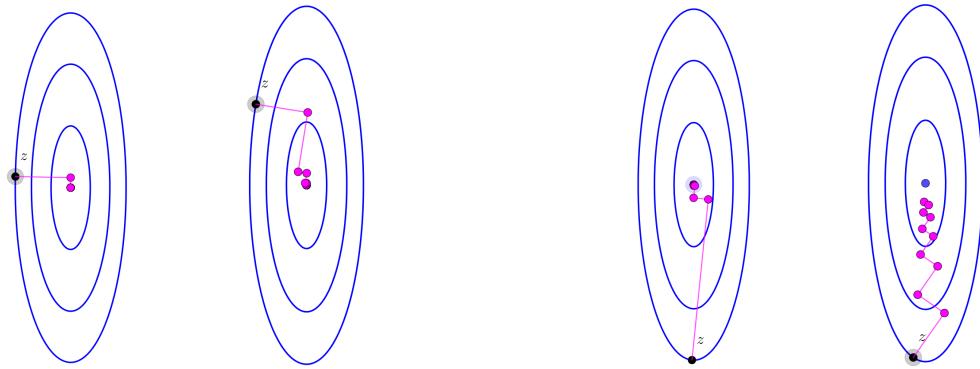


FIGURE IV.7 – Méthode du gradient optimal (GPO) pour une fonction mal conditionnée et des points initiaux perturbés.

Remarque IV.56 (Pas optimal et zig-zag). Comme on peut le voir sur la Figure IV.6, la méthode fonctionne mieux sur des fonctions bien conditionnées ; dans le cas contraire la méthode est ralentie par l’effet zig-zag. Comme on peut le voir également sur la Figure IV.7, l’effet zig-zag est également impacté par le choix du point initial. En particulier, on voit que lorsque on perturbe un peu un point initial situé dans l’espace propre de λ_{\max} , la trajectoire change peu, tandis que pour un point initial situé dans l’espace propre de λ_{\min} , la trajectoire est instable et très vite ralentie par les zig-zag.

Théorème IV.57 (Convergence de la méthode du gradient à pas optimal). Soit $f \in \Gamma_\mu(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$, pour $L > 0$. On considère la méthode du gradient à pas optimal. Alors, pour $\theta = \frac{L-\mu}{L+\mu}$, on a pour tout $k \in \mathbb{N}$:

$$f(x_{k+1}) - \inf f \leq \theta^2(f(x_k) - \inf f).$$

Démonstration. Admis. Une preuve est disponible en Annexe, dans la Section A.II.4. ■

Démonstration dans le cas quadratique. Cf. TD. ■

Remarque IV.58 (Adaptivité à la Lipschitzianité du gradient). Notez que l’on obtient exactement les mêmes vitesses que pour l’algorithme du gradient à pas fixe optimal (Théorèmes i) et IV.42) ! C’est d’autant plus remarquable qu’ici on ne définit aucun pas de temps ρ_k en fonction de L ou μ : on n’a pas besoin de connaître ces constantes pour l’algorithme fonctionne bien. On dit alors que la méthode du gradient à pas optimal est adaptive à la Lipschitzianité du gradient, au sens où elle n’a pas besoin de « savoir » que $f \in C_L^{1,1}(\mathbb{R}^N)$ pour bien fonctionner.

IV.IV Récapitulatif du Chapitre IV

On considère une fonction $f : \mathbb{R}^N \rightarrow \mathbb{R}$, et le problème d'optimisation associé

$$\text{minimiser}_{x \in \mathbb{R}^N} f(x).$$

Méthodes de descente

- $d \in \mathbb{R}^N$ est une direction de descente pour f en $x \in \mathbb{R}^N$ si $\langle \nabla f(x), d \rangle < 0$.
- Une méthode de descente est un algorithme de la forme

$$x_{k+1} = x_k + \rho_k d_k$$

où d_k est une direction de descente pour f en x_k .

- Une direction de descente de choix est $d_k = -\nabla f(x_k)$: cela donne la méthode du gradient.

Méthode du gradient à pas fixe (GPF)

- Si $f \in \Gamma_\mu(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$, et que l'on prend un pas fixe $\rho \in]0, 2/L[$, alors
 - 1) La suite des itérés $(x_k)_{k \in \mathbb{N}}$ générés par GPF converge vers un minimiseur x^* de f .
 - 2) La suite des valeurs $f(x_k)$ est décroissante et converge vers $\inf f$.
 - 3) Les itérés et les valeurs convergent linéairement : il existe $\theta \in [0, 1[$ tel que

$$\|x_{k+1} - x^*\| \leq \theta \|x_k - x^*\| \quad \text{et} \quad f(x_{k+1}) - \inf f \leq \theta^2 (f(x_k) - \inf f).$$

- 4) Le taux de convergence θ ne dépend que de ρ et du conditionnement de la fonction L/μ .
- 5) Le meilleur taux est obtenu lorsque $\rho = 2/(L + \mu)$.
- Si f n'est que convexe, la méthode marche encore mais elle converge moins vite.

Méthode du gradient à pas optimal (GPO)

- Ici on choisit ρ_k comme étant le pas qui fait décroître le plus possible la fonction :

$$\rho_k = \operatorname{argmin}_{\rho > 0} f(x_k - \rho \nabla f(x_k)).$$

- Les trajectoires générées zig-zaguent : $\langle \nabla f(x_k), \nabla f(x_{k+1}) \rangle = 0$.
- L'algorithme a exactement les mêmes propriétés de convergence que GPF avec le meilleur choix de pas $\rho = 2/(L + \mu)$.
- Pour une fonction quadratique, on peut calculer explicitement ρ_k sans avoir à connaître L ou μ .

Chapitre V

Optimisation sous contraintes

Dans ce chapitre nous nous intéressons aux problèmes d'optimisation **avec contrainte** :

$$(P_C) \quad \inf_{x \in C} f(x),$$

$f : U \rightarrow \mathbb{R}$, $C \subset U$ est non vide, où $U \subset \mathbb{R}^N$ est un ouvert. Jusqu'à présent nous avons plutôt ignoré la contrainte C :

- Dans le Chapitre II nous avons donné une Condition Nécessaire d'Optimalité lorsque \bar{x} est un minimiseur de f sur C qui se trouve être dans l'**intérieur** de la contrainte

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

Mais nous n'avons pas de CNO générale lorsque \bar{x} peut se trouver sur le bord de la contrainte. Or, en pratique, cette situation est la plus courante !

- Nous allons voir que de manière générale on peut décrire une CNO ayant la forme suivante :

$$\nabla f(\bar{x}) + \text{truc}(\bar{x}, C) = 0,$$

où $\text{truc}(\bar{x}, C)$ va être un nouvel objet dépendant de C et de \bar{x} , que l'on pourra interpréter comme « le gradient de C en \bar{x} », et qui bien sur s'annule lorsque $\bar{x} \in \text{int } C$.

Dans ce chapitre nous nous focaliserons sur le cas où la contrainte C peut s'écrire sous la forme d'équations et/ou inéquations.

V.I Introduction : Problèmes classiques

Pour des raisons historiques et pratiques, on tend à classer les problèmes d'optimisation sous contrainte en fonction de la nature la contrainte et de celle de f . Cette classification va, en gros, des problèmes les plus « simples » aux plus « compliqués¹ ».

¹En réalité c'est un peu plus complexe que cela mais on se limitera ici à cette présentation simplifiée.

V.I.1 Polyèdres

Définition V.1. On munit \mathbb{R}^M d'un ordre partiel dit canonique, noté \leq_M , défini par

$$x \leq_M y \Leftrightarrow (\forall i \in \{1, \dots, M\}) x_i \leq y_i.$$

Lorsqu'il n'y aura pas d'ambiguïté, on notera simplement \leq .

Remarque V.2. On manipule donc plusieurs relations d'ordre dans ce cours :

- L'ordre canonique dans \mathbb{R} : $1 \leq 2$
- L'ordre canonique dans \mathbb{R}^M : $(0, 2) \leq (1, 3)$
- L'ordre matriciel dans $\mathcal{M}_N(\mathbb{R})$: $A \succeq 0$.

Définition V.3 (Polyèdre). On dit que $C \subset \mathbb{R}^N$ est un **POLYÈDRE**² s'il existe $M \in \mathbb{N}$, $A \in \mathcal{M}_{M,N}(\mathbb{R})$ et $b \in \mathbb{R}^M$ tels que $C = [Ax \leq_M b]$.

Remarque V.4 (Polyèdre = Inégalités affines). On sait que une contrainte d'égalité linéaire de la forme $[Ax = b]$ décrit un sous-espace affine. Mais nous sommes moins familiers avec une contrainte d'inégalité affine $[Ax \leq_M b]$ telle qu'elle apparait dans la définition d'un polyèdre. A quoi ressemble cet ensemble ? Si on note $a_1, \dots, a_M \in \mathbb{R}^N$ tels que

$$A = \begin{pmatrix} a_1^\top \\ \vdots \\ a_M^\top \end{pmatrix},$$

on voit que la contrainte peut s'écrire comme l'intersection de M ensembles :

$$[Ax \leq b] = \{x \in \mathbb{R}^N \mid \forall i \in \{1, \dots, M\}, \langle a_i, x \rangle \leq b_i\} = \bigcap_{i=1}^M [\langle a_i, \cdot \rangle \leq b_i].$$

On sait que les solutions de $[\langle a_i, \cdot \rangle = b_i]$ constituent un hyperplan, porté par le vecteur a_i . On peut donc également facilement se convaincre (et c'est vrai) que $[\langle a_i, \cdot \rangle \leq b_i]$ est un demi-espace, délimité par l'hyperplan susmentionné (cf. Figures V.1). Donc un polyèdre, ce n'est rien d'autre qu'une intersection (finie) de demi-espaces.

²Prendre garde au fait que, dans la littérature française tant qu'anglophone, le terme *polyèdre* peut désigner des notions légèrement différentes. Il faut également faire attention à ne pas confondre avec *polygone* et *polytope*.

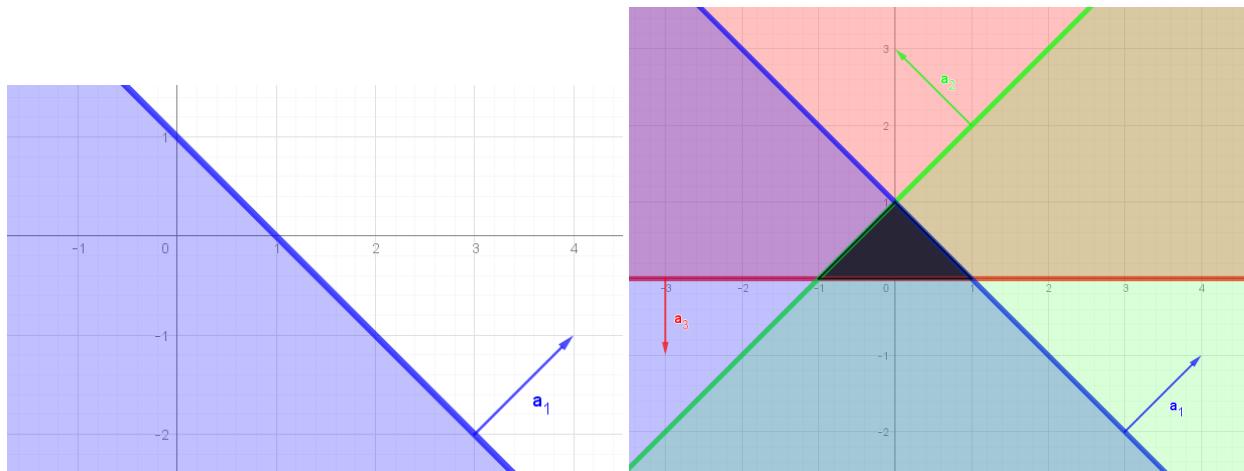


FIGURE V.1 – Gauche : En bleu, le demi-espace $\{z = (x, y) \in \mathbb{R}^2 \mid x + y \leq 1\}$, que l'on peut décrire comme $[\langle a_1, \cdot \rangle \leq b_1] = \{z = (x, y) \mid \langle a_1, z \rangle \leq b_1\}$ avec $a_1 = (1, 1)^\top$, $b_1 = 1$; En gras, l'hyperplan supporté par a_1 . Droite : Trois demi-espaces de la forme $[\langle a_i, \cdot \rangle \leq b_i]$ avec $a_1 = (1, 1)^\top$, $b_1 = 1$, $a_2 = (-1, 1)$, $b_2 = 1$ et $a_3 = (0, -1)$, $b_3 = 0$; et leur intersection, un triangle (en noir).

Exemple V.5 (Polyèdres). Faisons un peu de zoologie :

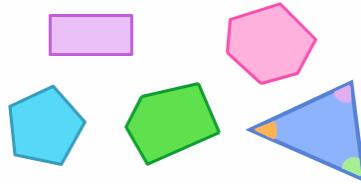


FIGURE V.2 – Quelques polyèdres bornés dans \mathbb{R}^2 . Ce sont des polygones convexes.

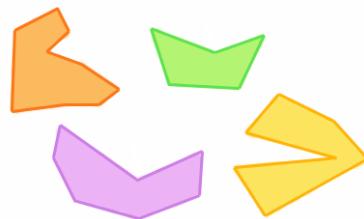


FIGURE V.3 – Ces polygones du plan ne sont pas des polyèdres.

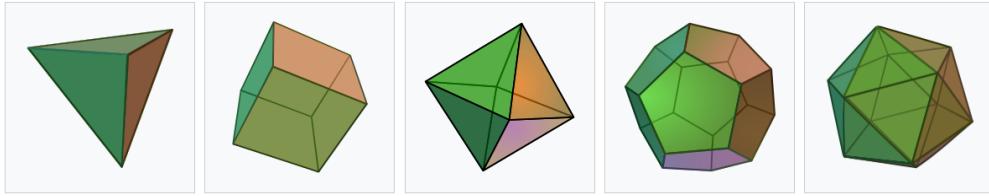


FIGURE V.4 – Cinq polyèdres bornés dans \mathbb{R}^3 (connus comme les cinq solides de Platon).

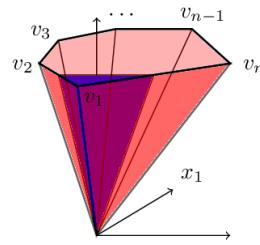


FIGURE V.5 – Un polyèdre de \mathbb{R}^3 qui est également un cône (non borné). Le cône a été tronqué afin de ne pas occuper un espace infini.

Remarque V.6 (Intersections de demi-espaces). Les polyèdres sont donc les ensembles que l'on obtient en intersectant un nombre *fini* de demi-espaces. On pourrait se demander ce qui se passe lorsque on prend une intersection *infinie* de demi-espaces ? La réponse est : cette procédure nous donne exactement tous les ensembles convexes ! C'est hors-programme, mais rien ne vous empêche de faire des dessins dans \mathbb{R}^2 pour vous en convaincre !

Exercice V.7 (Polyèdre et équation affine). Soient $A \in \mathcal{M}_{M,N}(\mathbb{R})$, $b \in \mathbb{R}^M$. Montrer que l'ensemble des solutions du problème linéaire associé

$$[Ax = b] = \{x \in \mathbb{R}^N \mid Ax = b\}$$

est un polyèdre.

Exercice V.8 (Polyèdre et espace affine). Montrer que tout sous-espace affine de \mathbb{R}^N est un polyèdre. On pourra commencer par le prouver pour un sous-espace vectoriel.

Exercice V.9 (Optimisation linéaire : Contrainte de boîte). Soient $\alpha, \beta \in \mathbb{R}^N$. Montrer que la boîte suivante :

$$C = \{x \in \mathbb{R}^N \mid \forall i = 1, \dots, M, \alpha_i \leq x_i \leq \beta_i\},$$

est un polyèdre.

Exercice V.10 (Polyèdre et convexité). Montrer que tout polyèdre est convexe.

V.I.2 Optimisation Linéaire

Les problèmes dits d'optimisation linéaire (*Linear Programming*, ou LP en VO) sont des problèmes d'optimisation où toutes les composantes sont **linéaires**. On cherche à minimiser une fonction linéaire sous une contrainte d'égalités ou inégalités affines³.

Définition V.11 (Optimisation linéaire). On dit qu'un problème d'optimisation est un problème d'**OPTIMISATION LINÉAIRE** s'il existe $A \in \mathcal{M}_{M,N}(\mathbb{R})$, $b \in \mathbb{R}^M$, $c \in \mathbb{R}^N$ tels que le problème s'écrive

$$\underset{x \in \mathbb{R}^N}{\text{minimiser}} \langle c, x \rangle \quad \text{tel que} \quad Ax \leq_M b. \quad (\text{V.1})$$

Exercice V.12 (Optimisation Linéaire : Contrainte de sous-niveaux). Soient $g_0, g_1, \dots, g_M : \mathbb{R}^N \rightarrow \mathbb{R}$ des fonctions affines. Montrer que le problème

$$\underset{x \in \mathbb{R}^N}{\text{minimiser}} g_0(x) \quad \text{tel que} \quad g_1(x) \leq 0, \dots, g_M(x) \leq 0,$$

est un problème d'optimisation linéaire.

Exemple V.13 (Le problème du transport optimal). Le problème du transport optimal consiste à trouver comment transporter, de la façon la plus efficace/économique possible, un objet d'un point A vers un point B . Ou, plus exactement, de nombreux objets depuis tout un tas de points de départ A_i vers des points d'arrivée B_i (cf. Figure V.6). Introduit à l'origine par [Monge](#) pour résoudre un problème de déplacement de tas de sable, ce problème permet de nos jours de répondre à des questions sur le « déplacement » d'objets plus abstraits, comme des images (cf. Figure V.7).

Ce problème peut être modélisé comme un problème d'optimisation, et plus précisément comme un problème de programmation linéaire. Pour plus de détails sur cette modélisation, vous pouvez lire [cet](#) et [cet](#) article.

³On pourrait se demander pourquoi on parle d'optimisation linéaire au lieu d'optimisation affine. Je pense que cela est du au fait que Kantorovich et Dantzig, fondateurs de la théorie, l'ont appelé ainsi et le nom est resté.

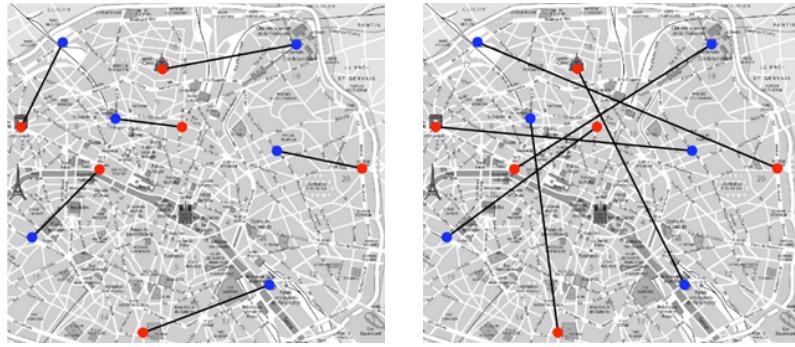


FIGURE V.6 – Si chaque point bleu doit aller sur un point rouge, lequel doit aller où pour minimiser la somme des trajets à vol d'oiseau? Et surtout : comment répondre à cette question sans avoir à tester les $n!$ combinaisons?



FIGURE V.7 – Application du Transport optimal : Une fois calculé un chemin optimal entre deux images (ici aux extrémités) on peut trouver au milieu de ce chemin une image (ici au centre) qui combine la forme d'une image avec le style de l'autre. Tout l'art ici consiste à définir correctement ce que « optimal » veut dire, qui est un problème *beaucoup* plus difficile que résoudre le problème de transport en lui-même. Extrait de l'article *Style transfer by relaxed optimal transport and self-similarity* par Kolkin et al., 2019 [11].

Exercice V.14 (Optimisation linéaire : forme standard). Soient $A \in \mathcal{M}_{M,N}(\mathbb{R})$, $b \in \mathbb{R}^M$, $c \in \mathbb{R}^N$. Montrer que le problème

$$\underset{x \in \mathbb{R}^N}{\text{minimiser}} \langle c, x \rangle \quad \text{tel que} \quad Ax = b \quad \text{et} \quad x \succeq_M 0, \quad (\text{V.2})$$

est équivalent à un problème d'optimisation linéaire. On appelle ce cas particulier un problème d'optimisation linéaire *sous forme standard*.

Exercice V.15 (Optimisation linéaire : canonique \Leftrightarrow standard). Soient $A \in \mathcal{M}_{M,N}(\mathbb{R})$, $b \in \mathbb{R}^M$, $c \in \mathbb{R}^N$ fixés. On considère les problèmes de programmation linéaire

- sous forme *canonique* associé à A, b, c : minimiser $\underset{x \in \mathbb{R}^N}{\langle c, x \rangle}$ tel que $Ax \leq b$,
- sous forme *standard* associé à A, b, c : minimiser $\underset{x \in \mathbb{R}^N}{\langle c, x \rangle}$ tel que $Ax = b$ et $x \succeq 0$.

- 1) Montrer que $\{x \in \mathbb{R}^N \mid Ax \leq b\} = \{x \in \mathbb{R}^N \mid \exists y \in \mathbb{R}^M : Ax + y = b \text{ et } y \succeq 0\}$.
- 2) Supposons que l'on veuille résoudre le problème sous forme *canonique* associé à A, b, c . Montrer qu'il existe un problème sous forme standard, dépendant de $\hat{A}, \hat{b}, \hat{c}$ (à trouver), tel que, si on le résolvait, nous donnerait immédiatement accès à la solution de notre problème sous forme canonique.
- 3) En déduire que les formes canonique et standard de l'optimisation linéaire sont équivalentes.

V.I.3 Optimisation Convexe

Définition V.16 (Optimisation convexe). On dit qu'un problème d'optimisation est un problème d'**OPTIMISATION CONVEXE** s'il existe $C \subset \mathbb{R}^N$ convexe et $f \in \Gamma_0(C)$ tels que le problème s'écrive

$$\underset{x \in C}{\text{minimiser}} f(x) \quad \text{tel que} \quad x \in C.$$

Le problème ci-dessus est dit sous forme *canonique*. Il est souvent bien pratique d'écrire un problème d'optimisation convexe sous sa forme *standard* :

Exercice V.17 (Optimisation convexe : forme standard). Soient $f, g_1, \dots, g_p : \mathbb{R}^n \rightarrow \mathbb{R}$ convexes et $h_1, \dots, h_q : \mathbb{R}^n \rightarrow \mathbb{R}$ affines.

- 1) Montrer que le problème

$$\underset{x \in \mathbb{R}^n}{\text{minimiser}} f(x) \quad \text{tel que} \quad \begin{cases} g_1(x) \leq 0, \dots, g_p(x) \leq 0 \\ h_1(x) = 0, \dots, h_q(x) = 0, \end{cases}$$

est un problème d'optimisation convexe.

- 2) Montrer que ce n'est pas forcément le cas si on suppose que les h_j sont convexes, en exhibant un contre-exemple.
- 3) Montrer que les problèmes d'optimisation linéaire sont convexes.

Exercice V.18 (Optimisation convexe : forme standard II). Montrer que pour tout problème d'optimisation convexe, il existe des fonctions $f, h : \mathbb{R}^N \rightarrow \mathbb{R}$ telles que le problème puisse se réécrire sous la forme

$$\underset{x \in \mathbb{R}^n}{\text{minimiser}} f(x) \quad \text{tel que} \quad h(x) = 0.$$

Même question avec

$$\underset{x \in \mathbb{R}^n}{\text{minimiser}} f(x) \quad \text{tel que} \quad g(x) \leq 0.$$

Remarque V.19 (Vocabulaire). On parlera parfois de

- problème d'optimisation convexe **sous contrainte d'égalité** pour désigner

$$\underset{x \in \mathbb{R}^n}{\text{minimiser}} f(x) \quad \text{tel que} \quad h_1(x) = 0, \dots, h_q(x) = 0,$$

- problème d'optimisation convexe **sous contrainte d'inégalité** pour désigner

$$\underset{x \in \mathbb{R}^n}{\text{minimiser}} f(x) \quad \text{tel que} \quad g_1(x) \leq 0, \dots, g_p(x) \leq 0,$$

- problème d'optimisation convexe **sous contrainte mixtes** pour désigner la forme standard

$$\underset{x \in \mathbb{R}^n}{\text{minimiser}} f(x) \quad \text{tel que} \quad \begin{cases} g_1(x) \leq 0, \dots, g_p(x) \leq 0 \\ h_1(x) = 0, \dots, h_q(x) = 0. \end{cases}$$

Au vu de l'Exercice V.18, il est légitime de se demander quel est l'intérêt de faire la différence entre toutes ces formes puisqu'elles sont équivalentes. D'une part, ce n'est pas parce que des problèmes sont équivalents qu'ils sont tous autant pratiques à résoudre. Prenons par exemple le cas des fonctions $\|x\| - 1$ et $\|x\|^2 - 1$ qui ont les mêmes sous-niveaux, mais dont l'une est différentiable et pas l'autre. D'autre part, d'un point de vue théorique, on verra qu'on aura besoin de vérifier des hypothèses, qui ne seront pas toujours vérifiées quelque soit la forme équivalente du problème.

Exemple V.20 (Problème de classification). On suppose que l'on dispose d'un certain type de données, et on veut être capable de les **classer** en deux groupes. Ce type de problème peut être très facile à réaliser pour un humain, mais toute la question est de savoir comment automatiser cette prise de décision pour l'implémenter sur une machine.

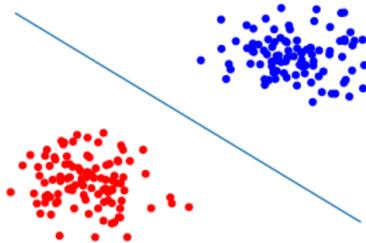


FIGURE V.8 – Classifier deux groupes de points dans \mathbb{R}^2 , relativement facile.

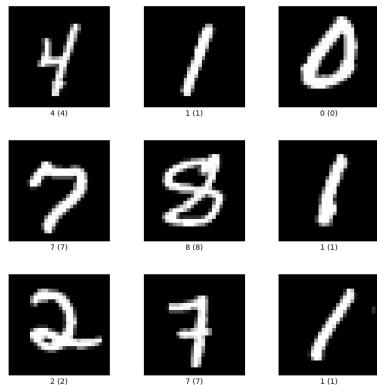


FIGURE V.9 – Classifier des nombres écrits à la main, difficulté moyenne. Issu du jeu de données [MNIST](#), utilisé abondamment pour tester les réseaux de neurones.

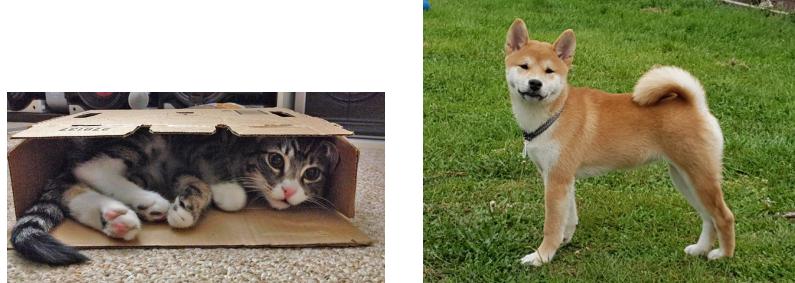


FIGURE V.10 – Classifier des photos dans \mathbb{R}^N , $N > 10^6$, en deux catégories (chat/chien), très difficile.

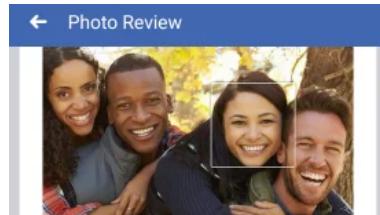


FIGURE V.11 – Classifier des visages humains, très très difficile.

Or, il est possible de modéliser ce problème en le transformant en un problème d'optimisation convexe, ayant la forme suivante :

$$\underset{x \in \mathbb{R}^N}{\text{minimiser}} \|x\|^2 \quad \text{tel que} \quad Ax \leq b,$$

où A et b sont construites à partir des données à classer. Dans ce contexte, ce problème est communément appelé **Machine à vecteur de support** (Support Vector Machine, ou SVM). Si le temps le permet, nous verrons comment modéliser et résoudre un tel problème (cf. feuille de TD5, et le TP associé).

V.II Théorème(s) de Lagrange-KKT pour l'optimisation sous contraintes d'égalités et inégalités

Dans cette section, on va s'intéresser aux problèmes s'écrivant sous la forme

$$\underset{x \in \mathbb{R}^n}{\text{minimiser}} f(x) \quad \text{tel que} \quad \begin{cases} g_1(x) \leq 0, \dots, g_p(x) \leq 0 \\ h_1(x) = 0, \dots, h_q(x) = 0, \end{cases}$$

où les fonctions en jeu seront convexes, affines ou quelconques, selon les besoins. Notre objectif est d'obtenir des Conditions d'Optimalité pour ces problèmes :

- Quel est l'équivalent de la CNO du 1er ordre que l'on avait dans le Théorème II.9 ? La réponse se trouve dans le Théorème V.34.
- Est-ce que cette CNO devient une CSO lorsque le problème est convexe, comme on l'avait vu dans le Théorème III.25 ? La réponse est : oui, voir le Théorème V.39.
- Est-ce que l'on peut avoir une CSO du 2e ordre, comme dans le Théorème II.19 ? Encore une fois, oui, cf. Théorème V.44.

Ces Théorèmes vont donc nous permettre de calculer à la main des minimiseurs locaux/globaux de problèmes d'optimisation sous contrainte, en résolvant des équations, de la même manière que l'on résolvait $\nabla f(x) = 0$ dans les premiers chapitres.

V.II.1 Contrainte d'inégalité simple et multiplicateur

On va commencer dans cette section par s'intéresser au problème simple où la contrainte s'écrit comme une contrainte d'inégalité. Autrement dit, lorsque

$$C = [g \leq 0],$$

pour $g : U \subset \mathbb{R}^N \rightarrow \mathbb{R}$ différentiable.

Proposition V.21. Soient $U \subset \mathbb{R}^N$ ouvert, $f, g : U \rightarrow \mathbb{R}$ et $C = [g \leq 0]$. Supposons que \bar{x} soit un minimiseur local de f sur C . Si $\nabla g(\bar{x}) \neq 0$, alors

$$g(\bar{x}) \leq 0 \quad \text{et} \quad (\exists \alpha \geq 0) \quad \nabla f(\bar{x}) + \alpha \nabla g(\bar{x}) = 0 \quad \text{et} \quad \alpha g(\bar{x}) = 0. \quad (\text{V.3})$$

Démonstration. Pour commencer, observons que le résultat est immédiat si $\nabla f(\bar{x}) = 0$, puisqu'il suffit de prendre $\alpha = 0$. Dans la preuve on supposera donc $\nabla f(\bar{x}) \neq 0$.

On va procéder en considérant deux cas : commençons par supposer que $g(\bar{x}) < 0$, on va voir qu'on aboutit à une contradiction. En effet, $\nabla f(\bar{x}) \neq 0$ implique via la Proposition IV.11 qu'il existe une direction de descente $d \in \mathbb{R}^N$ en \bar{x} , i.e. telle que $\langle \nabla f(\bar{x}), d \rangle < 0$. D'après le Lemme d'Armijo IV.12, cela veut dire que

$$(\exists \delta > 0)(\forall t \in]0, \delta[) \quad f(\bar{x} + td) < f(\bar{x}).$$

D'autre part, puisque g est continue et $g(\bar{x}) < 0$, on sait que pour t petit on aura encore $g(\bar{x} + td) < 0$. Autrement dit, $\bar{x} + td \in C$ et $f(\bar{x} + td) < f(\bar{x})$, ce qui contredit le fait que \bar{x} soit un minimiseur local. Ceci conclut la preuve dans le cas $g(\bar{x}) < 0$.

Supposons maintenant que $g(\bar{x}) = 0$. Dans un premier temps, nous allons montrer que $\nabla f(\bar{x}) \in \text{Vect}(\nabla g(\bar{x}))$. Raisonnons par l'absurde, et supposons que $\nabla f(\bar{x}) \notin \text{Vect}(\nabla g(\bar{x}))$. Puisque on a supposé que $\nabla g(\bar{x}) \neq 0$, cela veut dire que la famille $\{\nabla f(\bar{x}), \nabla g(\bar{x})\}$ est libre. Définissons la matrice dont les lignes sont ces gradients

$$A = \begin{pmatrix} \nabla f(\bar{x})^\top \\ \nabla g(\bar{x})^\top \end{pmatrix} \in \mathcal{M}_{2,N}(\mathbb{R}).$$

Ses lignes étant libres, nous en déduisons que A est surjective. Donc il existe un $d \in \mathbb{R}^N$ tel que $Ad = e$, où $e = (-1, -1)^\top$. Autrement dit, il existe un $d \in \mathbb{R}^N$ tel que

$$\langle \nabla f(\bar{x}), d \rangle = -1 \quad \text{et} \quad \langle \nabla g(\bar{x}), d \rangle = -1. \quad (\text{V.4})$$

On a donc une direction de descente commune pour ces fonctions ! D'après le Lemme d'Armijo IV.12 appliqué à f et g , cela veut dire qu'il existe un $\delta > 0$ commun tel que

$$(\forall t \in]0, \delta[) \quad f(\bar{x} + td) < f(\bar{x}) \quad \text{et} \quad g(\bar{x} + td) < g(\bar{x}) \leq 0.$$

Autrement dit, pour un tel choix de $t \in]0, \delta[$, on a $\bar{x} + td$ qui est toujours dans la contrainte $[g \leq 0]$ (puisque $g(\bar{x} + td) < 0$), mais qui est meilleur que \bar{x} au sens où $f(\bar{x} + td) < f(\bar{x})$. On se rend alors compte que ceci est en contradiction avec le fait que \bar{x} soit un minimiseur local de f sur C .

Nous avons donc montré par l'absurde que $\nabla f(\bar{x})$ et $\nabla g(\bar{x})$ sont colinéaires. Autrement dit, qu'il existe un $\alpha \in \mathbb{R}$ tel que

$$\nabla f(\bar{x}) + \alpha \nabla g(\bar{x}) = 0. \quad (\text{V.5})$$

Il ne nous reste donc plus qu'à prouver $\alpha \geq 0$. Encore une fois, raisonnons par l'absurde et supposons que $\alpha < 0$. Si on pose $d' = -\nabla f(\bar{x})$, on voit que

$$\langle \nabla f(\bar{x}), d' \rangle = -\|\nabla f(\bar{x})\|^2 < 0 \quad \text{et} \quad \langle \nabla g(\bar{x}), d' \rangle = \frac{-1}{\alpha} \langle \nabla f(\bar{x}), d' \rangle = \frac{1}{\alpha} \|\nabla f(\bar{x})\|^2 < 0.$$

On voit que l'on a encore une direction de descente d' commune pour f et g , ce qui va impliquer pour les mêmes raisons que précédemment, une contradiction. ■

Remarque V.22 (Vocabulaire). Il y a beaucoup de choses dans cette Proposition V.21. Il va être utile par la suite de bien nommer les ingrédients de ce résultat :

- La condition $\nabla g(\bar{x}) \neq 0$, qui est essentielle pour garantir le résultat, est appelée **condition de qualification** de la contrainte. On parle par exemple de contrainte qualifiée.
- La propriété $g(\bar{x}) \leq 0$ ne fait que traduire le fait que \bar{x} appartient à la contrainte $C = [g \leq 0]$. Autrement dit, que le vecteur \bar{x} est admissible (au sens où il ne viole pas la contrainte). C'est pour cela que l'on parle en général de condition d'**ADMISSIBILITÉ**.
- On distinguera souvent le fait que \bar{x} vérifie $g(\bar{x}) = 0$ ou $g(\bar{x}) < 0$. Lorsque $g(\bar{x}) = 0$, on dira que la contrainte $[g \leq 0]$ est **active** en \bar{x} , ce qui traduit que l'on est sur le bord du sous-niveau. Dans le cas où $g(\bar{x}) < 0$, on parlera de contrainte **inactive**.
- Le coefficient α que l'on voit apparaître est appelé le **multiplicateur de Lagrange** associé à la contrainte. On voit ici que α est positif; on verra d'autres contextes dans lequel le multiplicateur n'a pas de signe prescrit.
- La condition $\nabla f(\bar{x}) + \alpha \nabla g(\bar{x}) = 0$ est appelée la **condition de stationnarité** du problème. On vient de voir ici que c'est une condition *nécessaire* pour \bar{x} d'être un minimiseur local.
- La propriété $\alpha g(\bar{x}) = 0$ est la **condition de complémentarité** de la contrainte. Elle peut se reformuler de façon équivalente en :

$$\text{Si } g(\bar{x}) < 0 \text{ alors } \alpha = 0.$$

En d'autres termes, si la contrainte est inactive en \bar{x} , alors le multiplicateur de Lagrange est nul. Observer que dans ce cas la condition de stationnarité de Lagrange se réduit à $\nabla f(\bar{x}) = 0$. On voit que $\nabla g(\bar{x})$ a disparu de la condition de stationnarité, ce qui traduit le fait que la contrainte est inactive.

Remarque V.23 (Le système d'(in)équations de Lagrange-KKT). En pratique, lorsque on cherche un minimiseur de f sur $[g \leq 0]$, il faut donc chercher un couple $(x, \alpha) \in \mathbb{R}^N \times \mathbb{R}$ solution du système :

$$\begin{cases} \nabla f(x) + \alpha \nabla g(x) = 0 & (\text{Condition de stationnarité}) \\ g(x) \leq 0 & (\text{Condition d'admissibilité}) \\ \alpha \geq 0 & (\text{Multiplicateur}) \\ \alpha g(x) = 0 & (\text{Condition de complémentarité}) \end{cases}.$$

Une fois qu'on dispose de ces solutions, déterminer si elles sont des minimiseurs ou pas se fait exactement (aussi difficilement donc) comme on le fait pour les problèmes sans contraintes.

Exercice V.24 (Fonction quadratique sous contrainte d'inégalité linéaire). Soient $f(x, y) = \frac{1}{2}(x^2 + y^2) - 2x$ et $C = \{(x, y) \in \mathbb{R}^2 \mid x + y \leq 1\}$.

- 1) Montrer que f admet un unique minimiseur sur C .
- 2) Écrire les conditions d'optimalité pour ce problème, et trouver le minimiseur en résolvant le système associé.
- 3) La contrainte est elle active⁴ en cette solution ?

Exercice V.25 (Fonction quadratique sous contrainte d'inégalité linéaire II). Soient $f(x, y) = 2x - y$ et $C = \{(x, y) \in \mathbb{R}^2 \mid \frac{1}{2}x^2 + y^2 \leq 1\}$.

- 1) Montrer que f admet un minimiseur sur C .
- 2) Montrer que la contrainte est forcément qualifiée en ce minimiseur.
- 3) Écrire les conditions d'optimalité pour ce problème, les résoudre, et en déduire l'unique minimiseur de f sur C .
- 4) La contrainte est elle active en cette solution ?

Exercice V.26 (Problème non régulier). Soit $f(x) = -x^2$ et $g(x) = (|x| - 1)_+^2$.

- 1) Tracer le graphe de g , et calculer $C := [g \leq 0]$.
- 2) Tracer le graphe de f , et en déduire quels sont les minimiseurs de f sur C .
- 3) Vérifier que la condition d'optimalité de Lagrange-KKT n'est pas vérifiée en ces points, et expliquer pourquoi.

Exercice V.27 (Minimiser sur une boule). Soit $f : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiable, $a \in \mathbb{R}^N$ quelconque, et $C = \mathbb{B}(a, \delta)$ une boule fermée centrée en a de rayon $\delta > 0$. On suppose que \bar{x} est un minimiseur local de f sur C , et on va essayer d'écrire sa condition nécessaire d'optimalité.

- 1) Vérifier que $C = [g \leq 0]$, pour $g(x) = \|x - a\|^2 - \delta^2$, et calculer ∇g .
- 2) On suppose que la contrainte n'est pas qualifiée en \bar{x} (c-à-d. $\nabla g(\bar{x}) = 0$). Montrer que $\nabla f(\bar{x}) = 0$.
- 3) On suppose que la contrainte est qualifiée en \bar{x} (c-à-d. $\nabla g(\bar{x}) \neq 0$). Prouver que

$$\begin{cases} \text{si } \|x - a\| < \delta \text{ alors } \nabla f(x) = 0, \\ \text{si } \|x - a\| = \delta \text{ alors } (\exists \alpha \geq 0) \quad \nabla f(x) + \alpha(x - a) = 0. \end{cases}$$

- 4) En déduire qu'il existe $\alpha \geq 0$ tel que $\nabla f(\bar{x}) + \alpha(x - a) = 0$.

$$(\exists \alpha \geq 0) \quad \nabla f(x) + \alpha(x - a) = 0.$$

⁴On rappelle que pour une contrainte d'inégalité $[g \leq 0]$, la contrainte est dite active en x si $g(x) = 0$ (en d'autres termes on est sur le bord de la contrainte).

V.II.2 Condition d'Optimalité de KKT du 1er ordre

V.II.2.i) Introduction et définitions

On a vu dans la Proposition V.21 que pour minimiser une fonction f en présence d'une contrainte d'inégalité simple

$$g(x) \leq 0,$$

une condition nécessaire d'optimalité est (V.3), qui demande en particulier la condition de stationnarité

$$\nabla f(\bar{x}) + \alpha \nabla g(\bar{x}) = 0.$$

On peut donc se demander ce qui se passe lorsqu'on a affaire à *plusieurs* inégalités

$$g_1(x) \leq 0, \dots, g_p(x) \leq 0 ?$$

Ou à *plusieurs égalités*

$$h_1(x) = 0, \dots, h_q(x) = 0 ?$$

Ou à une combinaison des deux (on parle de contrainte *mixte*) :

$$C = \{x \in \mathbb{R}^N \mid g_1(x) \leq 0, \dots, g_p(x) \leq 0, h_1(x) = 0, \dots, h_q(x) = 0\}. \quad (\text{V.6})$$

En extrapolant un peu, il est raisonnable d'espérer que la condition de stationnarité devienne :

$$\nabla f(\bar{x}) + \alpha_1 \nabla g_1(\bar{x}) + \dots + \alpha_p \nabla g_p(\bar{x}) + \beta_1 \nabla h_1(\bar{x}) + \dots + \beta_q \nabla h_q(\bar{x}) = 0.$$

Comme nous allons le voir, cela est essentiellement vrai, les différences principales avec la Proposition V.21 étant que :

- les multiplicateurs β_j associés aux contraintes d'égalité n'ont pas de signe imposé,
- l'hypothèse de contrainte qualifiée ($\nabla g(\bar{x}) \neq 0$) va devenir un peu plus compliquée.

Avant d'énoncer notre premier Théorème V.34, donnons quelques définitions qui vont nous permettre d'exprimer une hypothèse de contrainte qualifiée.

Définition V.28. Soient $g_1, \dots, g_p, h_1, \dots, h_q : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiables, soit $C = \cap_i [g_i \leq 0] \cap \cap_j [h_j = 0]$ la contrainte mixte associée, et soit $x \in C$. On définit l'ensemble des **CONTRAINTES ACTIVES** en x par

$$I(x) = \{i \in \{1, \dots, p\} \mid g_i(x) = 0\}.$$

Remarque V.29 (Contraintes actives). Il faut noter que la notion de « contrainte active » ne vaut que pour les contraintes d'*inégalité*.

Définition V.30. Soient $g_1, \dots, g_p, h_1, \dots, h_q : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiables, soit $C = \cap_i [g_i \leq 0] \cap \cap_j [h_j = 0]$ la contrainte mixte associée, et soit $x \in C$. On dit que la contrainte mixte C est **QUALIFIÉE** en x si la famille de gradients

$$\{\nabla g_i(x), \nabla h_j(x)\}_{i \in I(x), 1 \leq j \leq q}$$

est linéairement indépendante.

Remarque V.31 (Contraintes actives 2). Si la famille de tous les vecteurs $\{\nabla g_i(x), \nabla h_j(x)\}_{1 \leq i \leq p, 1 \leq j \leq q}$ est libre, alors il n'y a pas besoin de calculer $I(x)$ puisque toute sous-famille sera également libre. Mais en pratique, il arrive souvent que $I(x)$ soit beaucoup plus petite que $\{1, \dots, p\}$, ce qui fait qu'il est plus facile ainsi de vérifier que les contraintes sont qualifiées.

Remarque V.32 (Qualification pour une unique contrainte). Si la contrainte est unique, alors la condition de qualification de la contrainte est drastiquement simplifiée :

- si on parle d'une contrainte d'égalité $[h = 0]$, que la famille $\{\nabla h(x)\}$ soit libre est équivalent à ce que $\nabla h(x) \neq 0$;
- si on parle d'une contrainte d'inégalité $[g \leq 0]$, une condition suffisante pour que la contrainte soit qualifiée est que $\nabla g(x) \neq 0$.

Noter que $\nabla g(x) \neq 0$ est exactement l'hypothèse de qualification que l'on a faite dans la Proposition V.21 !

Définition V.33 (Contrainte régulière). Soient $g_1, \dots, g_p, h_1, \dots, h_q : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiables, soit $C = \cap_i [g_i \leq 0] \cap \cap_j [h_j = 0]$ la contrainte mixte associée, et soit $x \in C$. On dit que la contrainte mixte C est **RÉGULIÈRE** en x si l'une des deux propriétés est vérifiée :

- toutes les fonctions $g_1, \dots, g_p, h_1, \dots, h_q$ sont affines ;
- la contrainte est qualifiée en x .

V.II.2.ii) Résultats principaux et commentaires

Nous sommes maintenant prêts à énoncer le premier Théorème de cette section, qui établit la **Condition Nécessaire d'Optimalité de KKT du 1er ordre** :

Théorème V.34 (Théorème de Lagrange-KKT : CNO du 1er ordre).

Soient $f, g_1, \dots, g_p, h_1, \dots, h_q : \mathbb{R}^N \rightarrow \mathbb{R}$ de classe C^1 . Soit $C = \cap_i [g_i \leq 0] \cap \cap_j [h_j = 0]$ la contrainte mixte associée. Supposons que \bar{x} soit un minimiseur local de f sur C . Si la contrainte

est régulière en \bar{x} , alors \bar{x} vérifie la Condition Nécessaire d'Optimalité de KKT du 1er ordre :

$$(\exists \alpha \in \mathbb{R}^p)(\exists \beta \in \mathbb{R}^q) \quad \begin{cases} \nabla f(\bar{x}) + \sum_{i=1}^p \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^q \beta_j \nabla h_j(\bar{x}) = 0 \\ \forall i = 1, \dots, p \quad g_i(\bar{x}) \leq 0 \\ \forall j = 1, \dots, q \quad h_j(\bar{x}) = 0 \\ \forall i = 1, \dots, p \quad \alpha_i \geq 0 \\ \forall i = 1, \dots, p \quad \alpha_i g_i(\bar{x}) = 0. \end{cases} \quad (\text{V.7})$$

Remarque V.35 (Point critique). On dira que \bar{x} est un **point critique** du problème si il vérifie la Condition Nécessaire d'Optimalité de KKT du 1er ordre. Le Théorème précédent nous dit donc que les points critiques sont de bons candidats à être des minimiseurs locaux.

Remarque V.36 (Le système d'(in)équations de KKT II). En pratique, lorsque on cherche un minimiseur de f sur une contrainte mixte, il faut donc chercher $(x, \alpha_1, \dots, \alpha_p, \beta_1, \dots, \beta_q) \in \mathbb{R}^N \times \mathbb{R}^p \times \mathbb{R}^q$ solution du système :

$$\begin{cases} \nabla f(x) + \sum_{i=1}^p \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^q \beta_j \nabla h_j(\bar{x}) = 0 & \text{(Condition de stationnarité)} \\ \forall i = 1, \dots, p \quad g_i(\bar{x}) \leq 0 & \text{(Condition d'admissibilité : inégalités)} \\ \forall j = 1, \dots, q \quad h_j(\bar{x}) = 0 & \text{(Condition d'admissibilité : égalités)} \\ \forall i = 1, \dots, p \quad \alpha_i \geq 0 & \text{(Multiplicateur : inégalités)} \\ \forall i = 1, \dots, p \quad \alpha_i g_i(\bar{x}) = 0 & \text{(Condition de complémentarité)} \end{cases}$$

Remarque V.37 (Historique et vocabulaire). Dans la littérature, ces conditions d'optimalités sont appelées conditions de Lagrange, ou parfois conditions de KKT (pour Karush-Kuhn-Tucker). Les raisons sont essentiellement historiques :

- **Joseph-Louis Lagrange** s'intéresse vers la fin du 18e siècle à des problèmes de mécanique, qui l'amènent à minimiser certaines quantités sous des contraintes d'égalité (voir Figure V.37). Il énonce alors une version du Théorème V.34 pour des contraintes d'égalité, introduisant l'idée de ces variables supplémentaires que l'on appelle désormais les multiplicateurs de Lagrange. On cite parfois ce résultat comme le *Théorème des multiplicateurs de Lagrange*, mais également comme le *Théorème des extrémas liés*.

3. De là résulte donc cette règle extrêmement simple pour trouver les conditions de l'équilibre d'un système quelconque proposé.

On prendra la somme des *moments* de toutes les puissances qui doivent être en équilibre (Sect. II, art. 5), et l'on y ajoutera les différentes fonctions différentielles qui doivent être nulles par les conditions du problème, après avoir multiplié chacune de ces fonctions par un coefficient indéterminé; on égalera le tout à zéro, et l'on aura ainsi une équation différentielle qu'on traitera comme une équation ordinaire de *maximis et minimis*, et d'où l'on tirera autant d'équations particulières finies qu'il y aura de variables. Ces équations étant ensuite débarrassées, par l'élimination, des coefficients indéterminés, donneront toutes les conditions nécessaires pour l'équilibre.

Cette équation donnera, relativement à chaque coordonnée, telle que x , de chacun des corps du système, une équation de la forme suivante

$$P \frac{\partial p}{\partial x} + Q \frac{\partial q}{\partial x} + R \frac{\partial r}{\partial x} + \dots + \lambda \frac{\partial L}{\partial x} + \mu \frac{\partial M}{\partial x} + \nu \frac{\partial N}{\partial x} + \dots = 0;$$

en sorte que le nombre de ces équations sera égal à celui de toutes les coordonnées des corps. Nous les appellerons *équations particulières de l'équilibre*.

FIGURE V.12 – Extrait du traité de *Mécanique Analytique* de Lagrange (1788) [13]. En français dans le texte.

- Au milieu du 20e siècle, la question de résoudre des problèmes d'optimisation sous contraintes générales d'inégalité se pose. En 1951, [Harold Kuhn](#) et [Albert Tucker](#)⁵ publient un article (intitulé *Nonlinear Programming* [12]) proposant des conditions d'optimalité pour ce problème. Cet article connaîtra un grand succès et aura beaucoup d'influence dans les décennies qui ont suivi, donnant naissance à un champ de recherche connu comme *l'optimisation non-linéaire*, et s'appliquant dans de nombreux domaines, allant de l'économie à l'ingénierie.

De manière surprenante, on se rendra compte près de 20 ans plus tard que ce résultat avait déjà été obtenu par [William Karush](#) dans... son mémoire de Master [9] datant de 1939 ! Depuis lors, les conditions d'optimalité (V.7) sont connues comme les conditions de Karush-Kuhn-Tucker, ou simplement KKT.

⁵Vous connaissez certainement déjà Tucker sans le savoir, puisqu'il est à l'origine du fameux « dilemne du prisonnier ». Il a beaucoup travaillé sur la Théorie des Jeux, et a notamment dirigé la thèse de John Nash sur ce sujet (1950), qui vaudra à ce dernier un prix Nobel en sciences économiques (1994).

First let me say that you have clear priority on the results known as the Kuhn–Tucker conditions (including the constraint qualification). I intend to set the record as straight as I can in my talk. [Kuhn, 1975a]

“you must be a saint” not to complain about the absence of recognition. [Kuhn, 1975a]

That does not answer the question of why I did not point to my work in later years when nonlinear programming took hold and flourished. The thought of doing this did occur to me from time to time, but I felt rather diffident about that early work and I don’t think I have a strong necessity to be “recognized.” In any case, the master’s thesis lay buried until a few years ago when Hestenes urged me to look at it again to see if it shouldn’t receive its proper place in history. . . So I did look at the thesis again, and I looked again at your work with Tucker. I concluded that you two had exploited and developed the subject so much further than I, that there was no justification for my announcing to the world, “Look what I did, first.” [Karush, 1975]

FIGURE V.13 – Extrait d’un échange de courrier entre Kuhn et Karush, dans lequel Kuhn s’engage à lui donner la reconnaissance qu’il mérite, et s’étonne que Karush ne se soit pas manifesté plus tôt [10].

Pour ces raisons, dans ce cours, nous parlerons toujours de conditions de KKT pour les problèmes d’optimisation sous contraintes mixtes.

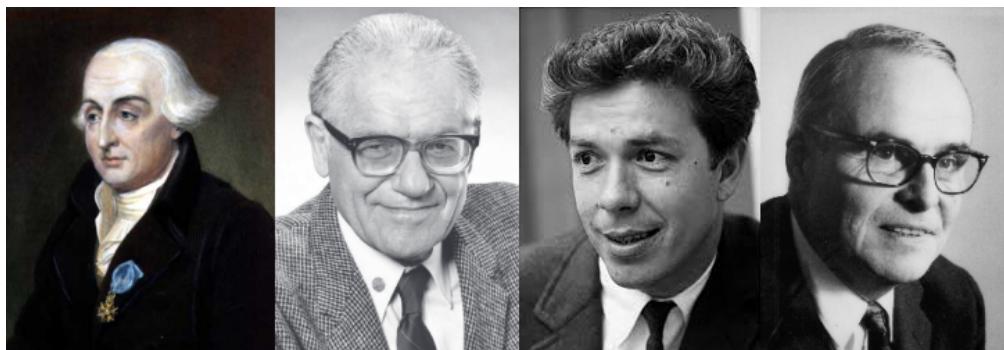


FIGURE V.14 – De gauche à droite : Lagrange, Karush, Kuhn et Tucker.

Remarque V.38 (Pourquoi les contraintes d’égalité ne se comportent pas comme les contraintes d’inégalité?). Si on regarde les conditions nécessaires d’optimalité de KKT, on voit qu’il y a une asymétrie entre les contraintes d’égalité et d’inégalité : les contraintes d’égalité n’ont pas

- de condition de compatibilité $\beta_j h_j(\bar{x}) = 0$,
- de condition sur les multiplicateurs $\beta_j \geq 0$.

Il est en fait assez facile de se convaincre qu'en fait ces deux conditions sont « trivialement » vérifiées, et n'ont donc pas lieu d'apparaître dans la condition nécessaire. En effet :

- la condition d'admissibilité $h_j(\bar{x}) = 0$ implique que $\beta_j h_j(\bar{x}) = 0$! Notez la différence avec les contraintes d'inégalité, pour lesquelles la condition d'admissibilité est $g_i(\bar{x}) \leq 0$, et pour laquelle on ne sait pas si $g_i(\bar{x}) = 0$. D'où l'importance pour ces dernières de faire la distinctions entre contraintes actives et non actives.
- on peut toujours forcer la contrainte $\beta_j \geq 0$ à être vérifiée, quitte à remplacer la fonction h_j par $-h_j$. Cela ne change rien au problème puisque $[h_j = 0] = [-h_j = 0]$. Notez la différence avec les contraintes d'inégalité, pour lesquelles on ne peut pas remplacer g_i par $-g_i$ sans changer la contrainte!

Voyons maintenant que cette CNO de KKT du 1er ordre est en fait une **CSO globale** lorsque le problème est convexe.

Théorème V.39 (Théorème de Lagrange-KKT : Réciproque convexe).

Soient $f, g_1, \dots, g_p, h_1, \dots, h_q : \mathbb{R}^N \rightarrow \mathbb{R}$ différentiables. Soit $C = \cap_i [g_i \leq 0] \cap \cap_j [h_j = 0]$ la contrainte mixte associée. Supposons que f, g_1, \dots, g_p soient convexes, et que h_1, \dots, h_q soient affines. Alors tout point $\bar{x} \in C$ qui vérifie la Condition Nécessaire d'Optimalité de KKT du 1er ordre (V.7) est un minimiseur global de f sur C .

On retrouve ainsi un analogue -sous contraintes- du Théorème III.25.

V.II.2.iii) Preuve des Théorèmes V.34 et V.39

Ici on considérera toujours que $f, g_1, \dots, g_p, h_1, \dots, h_q : \mathbb{R}^N \rightarrow \mathbb{R}$ sont de classe C^1 , et que $C = \cap_i [g_i \leq 0] \cap \cap_j [h_j = 0]$ est la contrainte mixte associée.

Lemme V.40 (de la pénalisation quadratique). Soit \bar{x} un minimiseur local de f sur C , et $I(\bar{x})$ les contraintes d'inégalités actives en \bar{x} . On considère pour tout $k \in \mathbb{N}$:

$$\phi_k(x) := f(x) + \frac{k}{2} \sum_{i \in I(\bar{x})} g_i(x)_+^2 + \frac{k}{2} \sum_{j=1}^q h_j(x)^2 + \frac{1}{2} \|x - \bar{x}\|^2.$$

Alors il existe une suite $x_k \rightarrow \bar{x}$, telle que $\nabla \phi_k(x_k) = 0$.

Démonstration. Soit \bar{x} un minimiseur local de f sur C . Par définition, il existe un $\varepsilon > 0$ tel que \bar{x} soit un minimiseur de f sur $\mathbb{B}(\bar{x}, \varepsilon)$. Il est clair que ϕ_k est une fonction continue, donc elle admet pour tout $k \in \mathbb{N}$ un minimiseur sur $\mathbb{B}(\bar{x}, \varepsilon)$, que l'on notera x_k . Notre objectif va être de montrer que $\lim_{k \rightarrow +\infty} x_k = \bar{x}$. Par définition cette suite est bornée, puisque contenue dans $\mathbb{B}(\bar{x}, \varepsilon)$. Il nous suffit donc de montrer que toute valeur d'adhérence de x_k est égale à \bar{x} .

Soit x_∞ une valeur d'adhérence de x_k . Commençons par vérifier que $x_\infty \in C$. D'une part, pour tout $i \notin I(\bar{x})$, on a $g_i(\bar{x}) < 0$. Donc, par continuité de g_i , et quitte à prendre ε plus petit, on a également $g_i(x_k) < 0$. D'autre part, l'optimalité de x_k nous permet d'écrire que

$$\begin{aligned}\phi_k(x_k) &\leq \phi_k(\bar{x}) \text{ car } x_k \text{ minimise } \phi_k \text{ sur } \mathbb{B}(\bar{x}, \varepsilon). \\ &= f(\bar{x}) \text{ car } g_i(\bar{x}) = 0, h_j(\bar{x}) = 0.\end{aligned}$$

Donc

$$f(\bar{x}) \geq \phi_k(x_k) \geq f(x_k) + \frac{k}{2} \sum_{i \in I(\bar{x})} g_i(x_k)_+^2 + \frac{k}{2} \sum_{j=1}^q h_j(x_k)^2.$$

Or $f(x_k)$ est minorée par $\inf_{\mathbb{B}(\bar{x}, \varepsilon)} f$, qui est indépendant de k . On voit donc que

$$0 \leq \frac{k}{2} \sum_{i \in I(\bar{x})} g_i(x_k)_+^2 + \frac{k}{2} \sum_{j=1}^q h_j(x_k)^2 \leq f(\bar{x}) - \inf_{\mathbb{B}(\bar{x}, \varepsilon)} f < +\infty.$$

Après division par k , on en déduit que les $g_i(x_k)_+^2$ et $h_j(x_k)^2$ tendent vers 0, ce qui implique que $g_i(x_\infty)_+^2 = 0$ et $h_j(x_\infty)^2 = 0$. Autrement dit, $g_i(x_\infty) \leq 0$ et $h_j(x_\infty) = 0$. On a donc bien montré que $x_\infty \in C$. Maintenant, on écrit

$$f(\bar{x}) \geq \phi_k(x_k) \geq f(x_k) + \frac{1}{2} \|x_k - \bar{x}\|^2,$$

et en passant à la limite on obtient

$$f(\bar{x}) \geq f(x_\infty) + \frac{1}{2} \|x_\infty - \bar{x}\|^2.$$

Or $x_\infty \in C \cap \mathbb{B}(\bar{x}, \varepsilon)$ et \bar{x} est un minimiseur local de f sur C . Donc $f(x_\infty) \geq f(\bar{x})$, et on en déduit que $x_\infty = \bar{x}$.

Maintenant qu'on sait que x_k tend vers \bar{x} , on peut dire que (à partir d'un certain rang) $x_k \in \text{int } \mathbb{B}(\bar{x}, \varepsilon)$. On peut donc appliquer le Théorème de Fermat II.10, qui nous dit dans ce cas que $\nabla \phi_k(x_k) = 0$. ■

Lemme V.41 (de Fritz John). Soit \bar{x} un minimiseur local de f sur C , et $I(\bar{x})$ les contraintes d'inégalités actives en \bar{x} . Alors

$$\lambda \nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^q \beta_j \nabla h_j(\bar{x}) = 0, \quad (\text{V.8})$$

où les multiplicateurs $\lambda \in \mathbb{R}_+$, $\alpha \in \mathbb{R}_+^{|I(\bar{x})|}$, $\beta \in \mathbb{R}^q$ sont non tous nuls.

Démonstration. Considérons le résultat du Lemme V.40 précédent. Après calcul du gradient, on obtient :

$$0 = \nabla \phi_k(x_k) = \nabla f(x_k) + \sum_{i \in I(\bar{x})} kg_i(x_k)_+ \nabla g_i(x_k) + \sum_{j=1}^q kh_j(x_k) \nabla h_j(x_k) + (x_k - \bar{x}).$$

Posons $\hat{\alpha}_{i,k} := kg_i(x_k)_+ \in \mathbb{R}_+$, $\hat{\beta}_{j,k} := kh_j(x_k) \in \mathbb{R}$. Alors :

$$0 = \nabla f(x_k) + \sum_{i \in I(\bar{x})} \hat{\alpha}_{i,k} \nabla g_i(x_k) + \sum_{j=1}^q \hat{\beta}_{j,k} \nabla h_j(x_k) + (x_k - \bar{x}). \quad (\text{V.9})$$

Considérons le vecteur réunissant les multiplicateurs $\hat{\pi}_k := (1, \hat{\alpha}_{i,k}, \hat{\beta}_{j,k}, 1)$. Alors $\|\hat{\pi}_k\|^2 = 1 + \sum \hat{\alpha}_{i,k}^2 + \sum \hat{\beta}_{j,k}^2 + 1$ est non nul. On peut donc définir $\pi_k := \hat{\pi}_k / \|\hat{\pi}_k\|$, constitué des coefficients $(\lambda_k, \alpha_{i,k}, \beta_{j,k}, \lambda_k)$, avec $\lambda_k = 1 / \|\hat{\pi}_k\|$, etc. Si on divise (V.9) par $\|\hat{\pi}_k\|$, on obtient donc

$$0 = \lambda_k \nabla f(x_k) + \sum_{i \in I(\bar{x})} \alpha_{i,k} \nabla g_i(x_k) + \sum_{j=1}^q \beta_{j,k} \nabla h_j(x_k) + \lambda_k (x_k - \bar{x}).$$

Maintenant, on observe que, par construction, $\|\pi_k\| = 1$, donc quitte à prendre une sous-suite, π_k converge vers un vecteur $\pi = (\lambda, \alpha_i, \beta_j, \lambda)$ de norme 1 lui aussi. Par ailleurs x_k converge vers \bar{x} , et les gradients sont continus. On peut donc passer à la limite et obtenir

$$0 = \lambda \nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^q \beta_j \nabla h_j(\bar{x}),$$

qui est exactement (V.8).

Pour conclure il nous faut vérifier quelques propriétés sur les multiplicateurs. D'une part, on a par définition que $\hat{\alpha}_{i,k} \geq 0$, donc $\alpha_{i,k} \geq 0$, et par passage à la limite $\alpha_i \geq 0$. De même, $\lambda_k = 1 / \|\hat{\pi}_k\| \geq 0$ donc $\lambda \geq 0$ aussi. D'autre part, on sait que $\pi = (\lambda, \alpha_i, \beta_j, \lambda)$ est de norme 1, donc non nul. D'où $(\lambda, \alpha, \beta) \neq 0$. ■

Lemme V.42 (Cas des contraintes qualifiées). *Considérons les hypothèses du Lemme V.41 de Fritz John. Supposons de plus que les contraintes sont qualifiées en \bar{x} . Alors $\lambda > 0$.*

Démonstration. On sait déjà d'après le Lemme V.41 que $\lambda \geq 0$. Supposons par l'absurde que $\lambda = 0$. Alors la condition d'optimalité (V.8) combinée avec $\lambda = 0$ veut dire que

$$\sum_{i \in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^q \beta_j \nabla h_j(\bar{x}) = 0.$$

Or la contrainte est qualifiée en \bar{x} , ce qui veut dire que la famille des gradients dans cette équation est libre. Le fait qu'on ait une combinaison linéaire nulle veut dire que l'on a forcément $\alpha_i = 0$ et $\beta_j = 0$. En d'autres termes $(\lambda, \alpha, \beta) = 0$. Ceci contredit le Lemme V.41 qui dit que les multiplicateurs (λ, α, β) sont non tous nuls. ■

Lemme V.43 (Cas des contraintes affines). Considérons les hypothèses du Lemme V.41 de Fritz John. Supposons de plus que les contraintes sont affines. Alors $\lambda > 0$.

Démonstration. Ici aussi, supposons par l'absurde que $\lambda = 0$. Considérons $x \in \mathbb{R}^N$ quelconque, et utilisons le fait que les contraintes soient affines pour écrire :

$$\begin{aligned} & \sum_{i \in I(\bar{x})} \alpha_i g_i(x) + \sum_{j=1}^q \beta_j h_j(x) \\ &= \sum_{i \in I(\bar{x})} \alpha_i g_i(\bar{x}) + \alpha_i \langle \nabla g_i(\bar{x}), x - \bar{x} \rangle + \sum_{j=1}^q \beta_j h_j(\bar{x}) + \beta_j \langle \nabla h_j(\bar{x}), x - \bar{x} \rangle \\ &= \left\langle \left(\sum_{i \in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^q \beta_j \nabla h_j(\bar{x}) \right), x - \bar{x} \right\rangle \\ &= 0, \end{aligned} \tag{V.10}$$

les deux dernières égalités venant du fait que $h_j(\bar{x}) = g_i(\bar{x}) = 0$, et du fait que $\lambda = 0$ dans (V.8). Nous allons maintenant montrer que la suite x_k introduite dans le Lemme V.40 viole cette égalité, ce qui nous permettra de conclure. Pour ce faire, nous allons revenir à comment cette suite et les multiplicateurs α_i, β_j ont été définis.

- Supposons qu'il existe $i \in I(\bar{x})$ tel que $\alpha_i \neq 0$. Alors $\alpha_i > 0$. Or α_i a été défini comme la limite de $\alpha_{i,k} = kg_i(x_k)_+ / \|\hat{\pi}_k\|$. Donc forcément, à partir d'un certain rang, $\alpha_{i,k} > 0$, ce qui implique que $g_i(x_k)_+ > 0$. Cette dernière inégalité est équivalente à dire que $g_i(x_k) > 0$. Nous en déduisons que $\alpha_i g_i(x_k) > 0$.
- Supposons qu'il existe j tel que $\beta_j \neq 0$. On a défini β_j comme la limite des $\beta_{j,k}$. Donc, à partir d'un certain rang, $\beta_{j,k}$ est non nul, et de même signe que β_j . Or $\beta_{j,k} = kh_j(x_k) / \|\hat{\pi}_k\|$. Donc, à partir d'un certain rang, $h_j(x_k)$ est non nul, et de même signe que β_j . On en déduit que $\beta_j h_j(x_k) > 0$.

On vient donc de montrer que si $(\alpha, \beta) \neq 0$ alors

$$\sum_{i \in I(\bar{x})} \alpha_i g_i(x_k) + \sum_{j=1}^q \beta_j h_j(x_k) > 0,$$

ce qui contredit (V.10). Cela veut donc dire que $(\alpha, \beta) = 0$. Or on a supposé que $\lambda = 0$, donc en fait $(\lambda, \alpha, \beta) = 0$, ce qui contredit le Lemme de Fritz John V.41. ■

Démonstration du Théorème V.34. Tout d'abord, observons que $\bar{x} \in C$ garantit déjà que $g_i(\bar{x}) \leq 0$ et $h_j(\bar{x}) = 0$. Ensuite, observons que la condition de complémentarité $\alpha_i g_i(\bar{x}) = 0$ est équivalente à dire que « $\alpha_i = 0$ ou $i \in I(\bar{x})$ ». Autrement dit, montrer (V.7) est équivalent à montrer que :

$$(\exists \alpha \in \mathbb{R}_+^p)(\exists \beta \in \mathbb{R}^q) \quad \nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^q \beta_j \nabla h_j(\bar{x}) = 0.$$

On fait appel au Lemme de Fritz John V.41 pour obtenir (V.8). On utilise ensuite le fait que la contrainte est régulière en \bar{x} , avec le Lemme V.42 ou V.43, pour obtenir que $\lambda > 0$. On peut alors diviser (V.8) par λ , et conclure. ■

Démonstration du Théorème V.39. On peut écrire

$$f(\bar{x}) \leq f(\bar{x}) + \sum_{i \in I(\bar{x})} \alpha_i(g_i(\bar{x}) - g_i(c)) + \sum_{j=1}^q \beta_j(h_j(\bar{x}) - h_j(c))$$

car $\alpha_i \geq 0$, $g_i(c) \leq 0$ par définition de C , $g_i(\bar{x}) = 0$ par définition de $I(\bar{x})$, $h_j(\bar{x}) = h_j(c) = 0$ par définition de C . Puisque on suppose les g_i convexes, on peut utiliser la caractérisation par les hyperplans tangents de la Proposition III.18 pour en déduire

$$f(\bar{x}) \leq f(\bar{x}) - \sum_{i \in I(\bar{x})} \alpha_i \langle \nabla g_i(\bar{x}), c - \bar{x} \rangle - \sum_{j=1}^q \beta_j(h_j(\bar{x}) - h_j(c)).$$

Puisque on suppose également les h_j affines, on peut également écrire

$$f(\bar{x}) \leq f(\bar{x}) - \sum_{i \in I(\bar{x})} \alpha_i \langle \nabla g_i(\bar{x}), c - \bar{x} \rangle - \sum_{j=1}^q \beta_j \langle \nabla h_j(\bar{x}), c - \bar{x} \rangle.$$

En utilisant maintenant la condition de KKT, avec la Proposition III.18 appliquée à f , on obtient

$$f(\bar{x}) \leq f(\bar{x}) + \langle \nabla f(\bar{x}), c - \bar{x} \rangle \leq f(c).$$

Ceci étant vrai pour tout $c \in C$, on conclut que \bar{x} est un minimiseur global de f sur C . ■

V.II.3 Condition d'Optimalité de KKT du 2e ordre

Passons maintenant à la **Condition Suffisante d'Optimalité de KKT du 2e ordre**, qui comme on se doute va faire intervenir une combinaison des hessiennes des contraintes :

Théorème V.44 (Théorème de Lagrange-KKT : CSO du 2e ordre).

Soient $f, g_1, \dots, g_p, h_1, \dots, h_q : \mathbb{R}^N \rightarrow \mathbb{R}$ deux fois différentiables. Soit $C = \cap_i [g_i \leq 0] \cap \cap_j [h_j = 0]$ la contrainte mixte associée. Supposons que \bar{x} vérifie :

- a) la Condition Nécessaire d'Optimalité de KKT du 1er ordre (V.7) avec des multiplicateurs $\bar{\alpha} \in \mathbb{R}^p$, $\bar{\beta} \in \mathbb{R}^q$;
- b) la définit positivité de la Hessienne Lagrangienne :

$$\nabla^2 f(\bar{x}) + \sum_{i=1}^p \bar{\alpha}_i \nabla^2 g_i(\bar{x}) + \sum_{j=1}^q \bar{\beta}_j \nabla^2 h_j(\bar{x}) \succ 0;$$

c) la condition de complémentarité stricte : $i \in I(\bar{x}) \Leftrightarrow \bar{\alpha}_i > 0$.

Alors \bar{x} est un minimiseur local de f sur C .

Remarque V.45 (Complémentarité stricte). Que veut dire cette complémentarité stricte ? Rappelons si nécessaire que dans la condition d'optimalité de KKT du 1er ordre, on demande une condition de complémentarité qui s'écrit

$$\alpha_i g_i(\bar{x}) = 0.$$

Comme on l'a déjà dit précédemment, ceci est équivalent à dire que

$$\alpha_i \neq 0 \Rightarrow g_i(\bar{x}) = 0.$$

Or, puisque on sait que $\alpha_i \geq 0$, et au vu de la définition de contrainte active, on voit que la condition de complémentarité est encore équivalente à

$$\alpha_i > 0 \Rightarrow i \in I(\bar{x}).$$

Cette complémentarité stricte demande donc un peu plus, à savoir *l'équivalence* entre ces deux propriétés.

Démonstration du Théorème V.44 sans inégalités. On commence par prouver ce résultat lorsqu'on a seulement des contraintes d'égalité. On introduit alors le Laplacien :

$$L(x) = f(x) + \sum_{j=1}^q \beta_j h_j(x),$$

qui vérifie $f = L$ sur C . Puisque \bar{x} vérifie la CNO de KKT du 1er ordre, on peut écrire :

$$\nabla L(\bar{x}) = \nabla f(\bar{x}) + \sum_{j=1}^q \beta_j \nabla h_j(\bar{x}) = 0.$$

De plus, b) nous donne :

$$\nabla^2 L(\bar{x}) = \nabla^2 f(\bar{x}) + \sum_{j=1}^q \beta_j \nabla^2 h_j(\bar{x}) \succ 0.$$

On voit alors que \bar{x} vérifie les conditions suffisantes d'optimalité du 2e ordre (sans contraintes) vues dans le Théorème II.19, ce qui implique que \bar{x} est un minimiseur local de L . Donc, pour tout $x \in C$ au voisinage de \bar{x} , on a

$$f(\bar{x}) = L(\bar{x}) \leq L(x) = f(x).$$

On en déduit donc que \bar{x} est un minimiseur local de f sur C . ■

Démonstration du Théorème V.44 : cas général. Maintenant passons au cas général avec des inégalités, et montrons qu'on peut se ramener au cas d'égalités seules. Quitte à réordonner les inégalités, et ce pour simplifier les notations, on va supposer que les premières correspondent aux contraintes actives. Autrement dit, $I(\bar{x}) = \{1, \dots, \bar{p}\}$ avec $\bar{p} \leq p$. On va définir un nouveau problème dans $\mathbb{R}^{N+\bar{p}}$: on introduit

$$\hat{f}(x, z) = f(x), \quad \hat{g}_i(x, z) = g_i(x) + z_i^2, \quad \hat{h}_j(x, z) = h_j(x), \quad \hat{C} = \bigcap_{i=1}^{\bar{p}} [\hat{g}_i = 0] \cap \bigcap_{j=1}^q [\hat{h}_j = 0].$$

On va s'intéresser au problème de minimiser \hat{f} sur \hat{C} . Notons que \hat{C} n'est défini que par des égalités ! De plus, il est facile de voir (en prenant $z_i = \sqrt{-g_i(x)}$) que

$$g_i(x) \leq 0 \text{ si et seulement si il existe } z_i \in \mathbb{R} \text{ tel que } \hat{g}_i(x, z_i) = 0.$$

On en déduit alors que x est un minimiseur local de f sur C si et seulement si il existe $z \in \mathbb{R}^{\bar{p}}$ tel que (x, z) soit minimiseur local de \hat{f} sur \hat{C} .

Considérons maintenant le \bar{x} de notre théorème, et définissons $\bar{z} \in \mathbb{R}^{\bar{p}}$ par $\bar{z}_i = \sqrt{-g_i(\bar{x})}$. Nous allons montrer que (\bar{x}, \bar{z}) est un minimiseur local de \hat{f} sur \hat{C} , ce qui terminera la preuve. Pour cela, il nous suffit de montrer que la condition suffisante du second ordre pour les contraintes d'égalités est vérifiée puisque on vient de le prouver ! On voit en particulier que si $i \in I(\bar{x})$ alors $\bar{z}_i = 0$. Grâce à notre hypothèse a), on peut écrire

$$\begin{aligned} & \nabla \hat{f}(\bar{x}, \bar{z}) + \sum_{i \in I(\bar{x})} \alpha_i \nabla \hat{g}_i(\bar{x}, \bar{z}) + \sum_{j=1}^q \beta_j \nabla \hat{h}_j(\bar{x}, \bar{z}) \\ &= \begin{pmatrix} \nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^q \beta_j \nabla h_j(\bar{x}) \\ \vdots \\ 2\alpha_i \bar{z}_i \\ \vdots \end{pmatrix} = \begin{pmatrix} 0_N \\ 0_{\bar{p}} \end{pmatrix}. \end{aligned}$$

On voit donc que (\bar{x}, \bar{z}) vérifie les conditions d'optimalité de KKT pour le problème de minimiser \hat{f} sur \hat{C} . On peut également écrire :

$$\begin{aligned} & \nabla^2 \hat{f}(\bar{x}, \bar{z}) + \sum_{i \in I(\bar{x})} \alpha_i \nabla^2 \hat{g}_i(\bar{x}, \bar{z}) + \sum_{j=1}^q \beta_j \nabla^2 \hat{h}_j(\bar{x}, \bar{z}) \\ &= \begin{pmatrix} \nabla^2 f(\bar{x}) + \sum_{i \in I(\bar{x})} \alpha_i \nabla^2 g_i(\bar{x}) + \sum_{j=1}^q \beta_j \nabla^2 h_j(\bar{x}) & 0 \\ 0 & 2 \text{Diag}(\alpha_i) \end{pmatrix} \end{aligned}$$

On a ici une matrice diagonale par blocs, dont le premier bloc est défini positif à cause de l'hypothèse b) ; et le deuxième bloc est la matrice diagonale $\text{Diag}(\alpha_i)$ qui est bien définie positive au vu de la condition de complémentarité stricte c). Donc cette grosse matrice

est bien définie positive. On voit donc que (\bar{x}, \bar{z}) vérifie la condition suffisante du second ordre pour le Théorème avec les contraintes d'égalité, que l'on a montré dans la première partie de la preuve. On en déduit donc que (\bar{x}, \bar{z}) est un minimiseur local de \hat{f} sur \hat{C} , ce qui implique que \bar{x} est un minimiseur local de f sur C . ■

Remarque V.46 (Sur une CNO de KKT du 2e ordre). Si on compare ces Théorèmes avec ceux que l'on a obtenus dans le cas sans contrainte, on voit qu'il nous en manque un : un analogue de la Condition Nécessaire d'Optimalité d'ordre 2 (cf. Théorème II.16). On s'attend à ce qu'il existe un résultat disant que : si \bar{x} est un minimiseur local de f sur C , et sous hypothèse que la contrainte soit régulière, alors non seulement la CNO de KKT du 1er ordre est satisfaite

$$\nabla f(\bar{x}) + \sum_{i=1}^p \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^q \beta_j \nabla h_j(\bar{x}) = 0,$$

mais de plus la combinaison de toutes ces Hessiennes sera semi-définie positive :

$$\nabla^2 f(\bar{x}) + \sum_{i=1}^p \alpha_i \nabla^2 g_i(\bar{x}) + \sum_{j=1}^q \beta_j \nabla^2 h_j(\bar{x}) \succeq 0. \quad (\text{V.11})$$

Le problème est qu'un tel résultat n'existe pas, malheureusement. Plus précisément :

- On peut trouver un contre-exemple avec un point qui est minimiseur local mais pour lequel la matrice dans (V.11) n'est pas semi-définie positive (voir Exemple V.47 suivant).
- On peut montrer un résultat un peu plus faible que (V.11), mais qui n'est pas vraiment facile à utiliser en pratique : « la matrice dans (V.11) est semi-définie positive dans les directions tangentes à la contrainte ». On ne s'étendra pas sur ce que cela veut dire, car cela dépasse le programme de ce cours.

Exemple V.47 (Un contre-exemple à l'existence d'une CNO de KKT 2e du ordre). Soit $f(x, y) = x^2 + y^2$ et $C = [h = 0]$ avec $h(x, y) = y$.

- 1) On a $C = \{(x, y) \in \mathbb{R}^2 \mid y = 0\}$, ce qui nous permet de voir que sur la contrainte, $f(x, y) = x^2$. On en déduit donc immédiatement que f admet un unique minimiseur sur C , qui est $(x, y) = (0, 0)$.
- 2) On voit que la contrainte est qualifiée en $(0, 0)$, puisque $\nabla g(0, 0) = (0, 1)^\top \neq (0, 0)^\top$. Donc la CNO de KKT du 1er ordre s'applique, et on obtient que $\nabla f(0, 0) + \beta \nabla g(0, 0) = 0$, pour un certain $\beta \in \mathbb{R}$. Puisque $\nabla f(0, 0) = (0, 0)^\top$ et $\nabla g(0, 0) = (0, 1)^\top$, on voit immédiatement que le multiplicateur β est nul.
- 3) On peut calculer

$$\nabla^2 f(0, 0) + \beta \nabla^2 g(0, 0) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix},$$

et on se rend compte que cette matrice n'est pas semi-définie positive.

On voit donc bien qu'une condition telle que (V.11) n'est pas vraie en général.

Exercice V.48 (Utilisation de la CSO de KKT du 2e ordre). Soient $f(x, y) = -x$, et $C = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1, (x - 1)^3 - y = 0\}$.

- 1) En utilisant la Condition Suffisante d'Optimalité du 2e ordre de KKT, montrer que $(1, 0)$ est un minimiseur local de f sur C .
- 2) Vérifier que $f(x, y) \geq -1$ pour tout $(x, y) \in C$. En déduire que $(1, 0)$ est un minimiseur de f sur C .
- 3) *Optionnel* : Dessinez C dans le plan, et convainquez-vous graphiquement que $(1, 0)$ est l'unique minimiseur de f sur C .

V.III Algorithmes pour l'optimisation sous contraintes

V.III.1 Projection sur un convexe fermé

Définition V.49. Soit $C \subset \mathbb{R}^N$ un ensemble non vide, et $x \in \mathbb{R}^N$. On définit la **PROJECTION** de x sur C comme étant le sous-ensemble de C (possiblement vide) défini par :

$$\text{proj}_C(x) := \underset{c \in C}{\operatorname{argmin}} \text{ dist}(c, x).$$

Remarque V.50 (Points fixes de la projection). Observer que x appartient à C si et seulement si $\text{proj}_C(x) = x$.

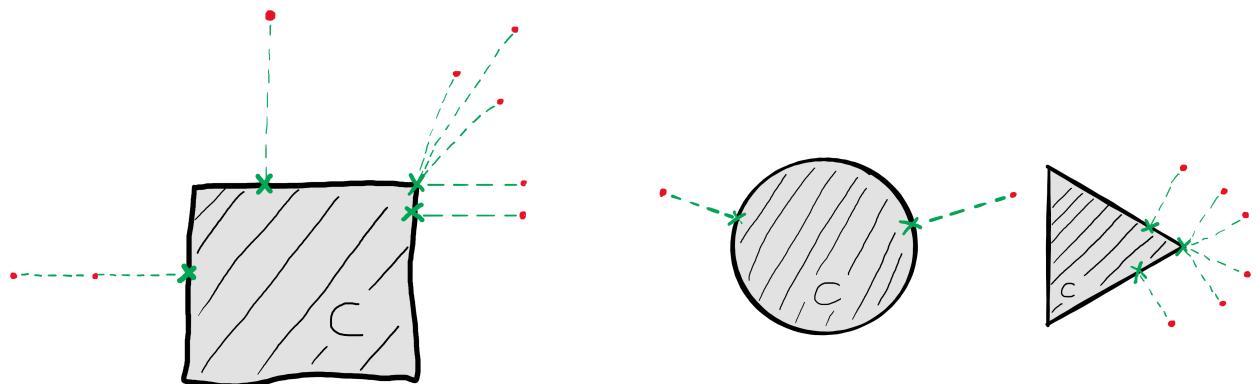


FIGURE V.15 – Diverses projections sur un carré. Des points différents (en rouge) peuvent se projeter sur le même point (en vert).

FIGURE V.16 – Encore quelques projections sur des convexes.

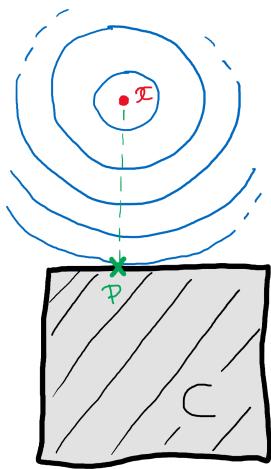


FIGURE V.17 – Un ensemble convexe C , un point x (en rouge) et sa projection $p = \text{proj}_C(x)$ sur C (en vert), qui est le point de C qui est le plus proche possible de x . Pour trouver cette projection on peut imaginer une boule centrée en x dont le rayon grossit jusqu'à toucher C : lorsque l'intersection entre cette boule et C est réduite à un point, alors ce point est exactement $\text{proj}_C(x)$.

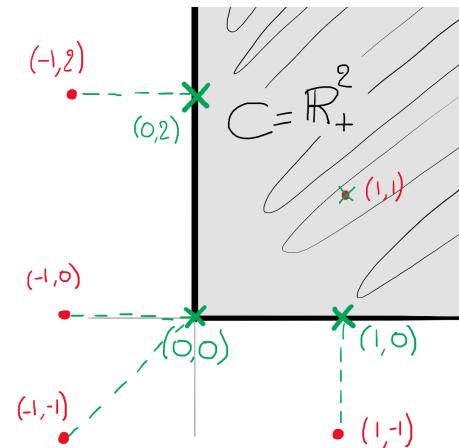


FIGURE V.18 – Divers points x (en rouge) et leurs projections $p = \text{proj}_C(x)$ (en vert) sur l'orthant positif $C = \mathbb{R}^2_+$. Dans ce cas la projection a pour effet de mettre tous les coefficients négatifs à zéro.

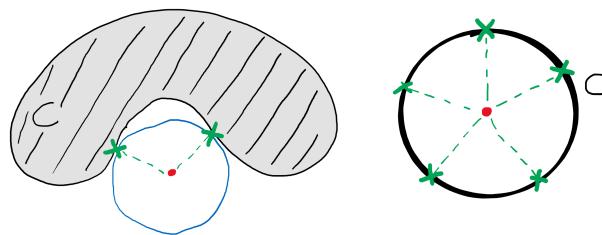


FIGURE V.19 – La projection n'est pas bien définie si C n'est pas convexe! Ici deux ensembles C non convexes, une patate et un cercle ($\text{cercle} \neq \text{disque}$) pour lesquels le point rouge peut trouver plus d'un point vert dans C qui minimise la distance.

Lorsque C est convexe fermé, la fonction $\text{proj}_C : \mathbb{R}^N \rightarrow C$ est bien définie :

Proposition V.51 (La projection est bien définie sur les convexes fermés). Soit $C \subset \mathbb{R}^N$ un ensemble non vide.

- i) Si C est fermé, alors $\text{proj}_C(x)$ est non vide pour tout $x \in \mathbb{R}^N$.
- ii) Si C est fermé et convexe, alors $\text{proj}_C(x)$ est réduit à exactement un unique point, pour tout $x \in \mathbb{R}^N$.

Démonstration. On écrit

$$\text{proj}_C(x) = \underset{c \in C}{\operatorname{argmin}} \text{dist}(c, x) = \underset{c \in C}{\operatorname{argmin}} \|c - x\|^2.$$

On observe que $x' \mapsto \|x' - x\|^2$ est fortement convexe et continue sur \mathbb{R}^N . D'après le Théorème III.38, on sait que $x' \mapsto \|x' - x\|^2$ est coercive sur \mathbb{R}^N .

- i) Si on suppose que C est fermé, alors $\text{proj}_C(x)$ est l'ensemble des minimiseurs d'une fonction continue coercive sur un fermé. D'après le Théorème II.35, on sait que cet ensemble de minimiseurs est non vide.
- ii) Si on suppose de plus que C est convexe, alors on peut dire que $x' \mapsto \|x' - x\|^2$ est fortement convexe sur C . Donc d'après le Théorème III.38, on sait qu'il y a exactement un minimiseur. ■

Exercice V.52. Calculer l'opérateur de projection pour les ensembles suivants :

- 1) $C = \{x \in \mathbb{R}^N \mid \|x\| \leq 1\}$
- 2) $C = C_1 \times \cdots \times C_N \subset \mathbb{R}^N$, où $C_1, \dots, C_N \subset \mathbb{R}$.
- 3) $C = \mathbb{R}_+^N = \{x = (x_1, \dots, x_N) \in \mathbb{R}^N \mid x_i \geq 0\}$, parfois appelé l'orthant positif.
- 4) $C = \{(x, y) \in \mathbb{R}^2 \mid y = 0\}$.

Le point projeté p de x sur C peut également se caractériser comme étant l'unique point tel que le vecteur $x - p$ forme un angle obtus⁶ avec tous les vecteurs entrants $c - p$, pour $c \in C$ (cf. Figures V.20 et V.21) :

Proposition V.53 (Caractérisation de la projection via les angles). Soit $C \subset \mathbb{R}^N$ un ensemble convexe fermé non vide. Soit $x \in \mathbb{R}^N$ et $p \in C$. Alors $p = \text{proj}_C(x)$ si et seulement si

$$(\forall c \in C) \quad \langle c - p, x - p \rangle \leq 0. \tag{V.12}$$

⁶On rappelle que deux vecteurs x et y forment un angle obtus si et seulement si $\langle x, y \rangle \leq 0$.

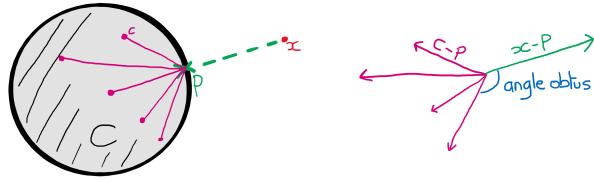


FIGURE V.20 – Caractérisation de la projection par les angles : on voit que si $p = \text{proj}_C(x)$, alors pour tout $c \in C$, le vecteur $c - p$ forme un angle obtus avec $x - p$.

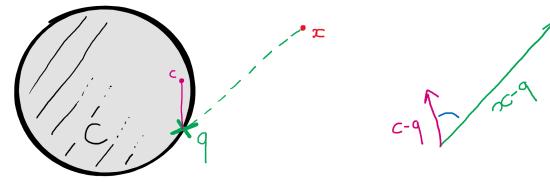


FIGURE V.21 – Caractérisation de la projection par les angles : on voit que si $q \notin \text{proj}_C(x)$, alors il existe un $c \in C$ tel que le vecteur $c - q$ forme un angle aigu avec $x - q$.

Démonstration. On va faire la preuve en deux temps. Supposons que $p = \text{proj}_C(x)$ et montrons que (V.12) est vérifiée. On se donne donc $c \in C$ quelconque, et pour tout $\alpha \in]0, 1[$ on considère $(1 - \alpha)p + \alpha c$ qui appartient à C par convexité. Alors, par définition de la projection,

$$\|x - p\|^2 \leq \|x - (1 - \alpha)p + \alpha c\|^2 = \|x - p\|^2 + \alpha^2\|c - p\|^2 + 2\alpha\langle x - p, p - c \rangle.$$

En simplifiant et en divisant par $\alpha > 0$, on obtient

$$0 \leq \alpha\|c - p\|^2 + 2\langle x - p, p - c \rangle.$$

Puisque ceci est vrai pour tout $\alpha \in]0, 1[$, on peut faire tendre $\alpha \rightarrow 0$, ce qui nous donne bien

$$0 \leq \langle x - p, p - c \rangle.$$

Supposons maintenant que $p \in C$ est un vecteur vérifiant (V.12), et montrons que c'est $\text{proj}_C(x)$. Par hypothèse, on a pour tout $c \in C$:

$$0 \geq \langle x - p, c - p \rangle = \langle x - p, c - x + x - p \rangle = \langle x - p, c - x \rangle + \|x - p\|^2.$$

En utilisant l'inégalité de Cauchy-Schwarz, et en divisant par $\|x - p\|^2$ (on peut le faire sauf si $p = x$ mais dans ce cas $p = \text{proj}_C(x)$ est trivial) on obtient :

$$0 \geq -\|x - p\|\|c - x\| + \|x - p\|^2 \Rightarrow 0 \geq \|x - p\| - \|c - x\|.$$

Ceci étant vrai pour tout $c \in C$, on en déduit que p est la projection de x sur C . ■

Exercice V.54. Soit C l'hyperplan affine défini par $C = \{x \in \mathbb{R}^N \mid \langle a, x \rangle = b\}$. Vérifier, à l'aide de la caractérisation de la projection via les angles, que :

$$\text{proj}_C(x) = x - \frac{\langle a, x \rangle - b}{\|a\|^2} a$$

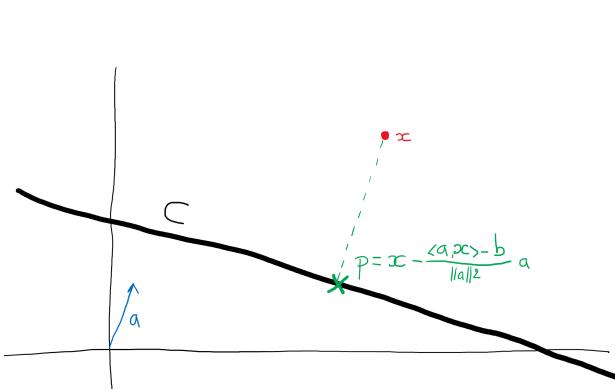


FIGURE V.22 – Projection sur une droite affine portée par a .

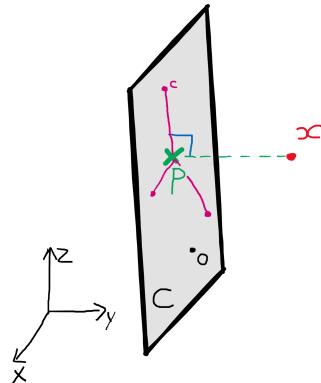


FIGURE V.23 – Projection sur un sous-espace vectoriel (ici un hyperplan). On peut voir que les vecteurs $x - p$ et $c - p$ forment un angle droit.

On déduit de la Proposition V.53 un corollaire sur la projection sur un espace vectoriel, caractérisée par le fait que $x - p$ doit former un angle *droit* avec tous les vecteurs de l'espace (Figure V.23) :

Corollaire V.55. Soit F un sous-espace vectoriel non vide de \mathbb{R}^N . Soient $x \in \mathbb{R}^N$ et $p \in F$. Alors $p = \text{proj}_F(x)$ si et seulement si

$$(\forall c \in F) \quad \langle c, x - p \rangle = 0.$$

Démonstration. On vient de voir que $p = \text{proj}_F(x)$ si et seulement si

$$(\forall c \in F) \quad \langle c - p, x - p \rangle \leq 0.$$

Or $p \in F$ donc par linéarité, $c \in F \Leftrightarrow c - p \in F$. Donc l'inégalité ci-dessus devient

$$(\forall c \in F) \quad \langle c, x - p \rangle \leq 0.$$

Mais de plus, $c \in F \Leftrightarrow -c \in F$ par linéarité, donc cette inégalité devient égalité :

$$(\forall c \in F) \quad \langle c, x - p \rangle = 0. \quad \blacksquare$$

Exercice V.56. Soit F un sous-espace vectoriel de \mathbb{R}^N non vide, et soit $p = \text{proj}_F$. Montrer que p est une application linéaire, et que p est la projection orthogonale sur F , au sens où :

$$p \circ p = p \quad \text{et} \quad \|p\| \leq 1.$$

V.III.2 Propriétés avancées de la projection

Les algorithmes pour résoudre des problèmes d'optimisation sous contrainte comportent souvent des projections à réaliser sur la contrainte. Nous allons donc avoir besoin de quelques propriétés sur la projection.

Lemme V.57 (Non-expansion ferme de la projection). Soit $C \subset \mathbb{R}^N$ convexe fermé non vide. Alors la projection $\text{proj}_C : \mathbb{R}^N \rightarrow \mathbb{R}^N$ est fermement non-expansive :

$$(\forall x, y \in \mathbb{R}^N) \quad \|\text{proj}_C(y) - \text{proj}_C(x)\|^2 \leq \|y - x\|^2 - \|(y - x) - (\text{proj}_C(y) - \text{proj}_C(x))\|^2.$$

Démonstration. (Voir [8, Proposition III.3.1.3]) Commençons par développer la norme au carré, en faisant apparaître les termes de projection :

$$\begin{aligned} & \|y - x\|^2 \\ &= \|(y - x) - (\text{proj}_C(y) - \text{proj}_C(x)) + (\text{proj}_C(y) - \text{proj}_C(x))\|^2 \\ &= \|(y - x) - (\text{proj}_C(y) - \text{proj}_C(x))\|^2 + \|\text{proj}_C(y) - \text{proj}_C(x)\|^2 \\ &\quad + 2\langle (y - x) - (\text{proj}_C(y) - \text{proj}_C(x)), \text{proj}_C(y) - \text{proj}_C(x) \rangle. \end{aligned}$$

On voit que le Lemme sera prouvé pourvu qu'on arrive à monter que le produit scalaire est positif. Coupons ce terme en deux :

$$\begin{aligned} & \langle (y - x) - (\text{proj}_C(y) - \text{proj}_C(x)), \text{proj}_C(y) - \text{proj}_C(x) \rangle \\ &= -\langle y - \text{proj}_C(y), \text{proj}_C(x) - \text{proj}_C(y) \rangle - \langle x - \text{proj}_C(x), \text{proj}_C(y) - \text{proj}_C(x) \rangle. \end{aligned}$$

On voit alors que chacun de ces deux produits scalaires est négatif, grâce à la caractérisation de la projection par les angles. D'où le résultat. ■

Théorème V.58 (La projection est 1-Lipschitzienne). Soit $C \subset \mathbb{R}^N$ convexe fermé non vide. Alors la projection $\text{proj}_C : \mathbb{R}^N \rightarrow \mathbb{R}^N$ est 1-Lipschitzienne (on dit aussi non-expansive) :

$$(\forall x, y \in \mathbb{R}^N) \quad \|\text{proj}_C(y) - \text{proj}_C(x)\| \leq \|y - x\|.$$

Démonstration. C'est une conséquence directe du Lemme de non-expansivité ferme V.57, où on élimine le terme négatif du second membre et on enlève les carrés. ■

Remarque V.59 (Contraction des distances). Cela veut dire que si on prend deux points puis qu'on les projette, les projections seront plus rapprochées que ne l'étaient les points de départ. On peut bien voir ce phénomène sur les Figures V.15 et V.16.

On termine avec un résultat montrant que la projection est liée à la dérivée de la fonction distance.

Proposition V.60 (Gradient de la distance au carré). Soit C un ensemble convexe fermé non vide, et $f : \mathbb{R}^N \rightarrow \mathbb{R}$ définie par $f(x) = \frac{1}{2} \text{dist}(x, C)^2$. Alors $f \in C_1^{1,1}(\mathbb{R}^N)$, avec

$$\nabla f(x) = x - \text{proj}_C(x).$$

Démonstration. Soient $x, y \in \mathbb{R}^N$ quelconques. Dans cette preuve on notera $p_x := \text{proj}_C(x)$ et $p_y := \text{proj}_C(y)$. On définira aussi la fonction $A : \mathbb{R}^N \rightarrow \mathbb{R}^N$, telle que $Ax := x - \text{proj}_C(x)$. Observons déjà que $A = I - \text{proj}_C$ est 1-Lipschitzienne, d'après le Lemme V.57 de non-expansion ferme de la projection. Notre objectif est maintenant de prouver que $Ax = \nabla f(x)$, en vérifiant la formule de Taylor :

$$f(y) - f(x) - \langle Ax, y - x \rangle = o(\|y - x\|). \quad (\text{V.13})$$

Commençons par montrer que

$$f(y) - f(x) - \langle Ax, y - x \rangle \geq 0. \quad (\text{V.14})$$

Pour cela, on observe que la définition de la projection nous permet d'écrire que $f(x) = (1/2)\|x - p_x\|^2 = (1/2)\|Ax\|^2$. On peut alors écrire

$$\begin{aligned} & f(y) - f(x) - \langle Ax, y - x \rangle \\ &= \frac{1}{2}\|Ay\|^2 - \frac{1}{2}\|Ax\|^2 - \langle Ax, y - x \rangle \\ &= \frac{1}{2}\|Ay\|^2 - \frac{1}{2}\|Ax\|^2 - \langle Ax, Ay - Ax \rangle - \langle Ax, p_y - p_x \rangle \text{ car } x = Ax + p_x \\ &= \frac{1}{2}\|Ay - Ax\|^2 - \langle x - p_x, p_y - p_x \rangle \text{ en réorganisant les termes.} \end{aligned}$$

Or ici on a $\|Ay - Ax\|^2 \geq 0$, et d'autre part via la caractérisation de la projection par les angles (Proposition V.53), on a $\langle x - p_x, p_y - p_x \rangle \leq 0$. On a donc bien prouvé (V.14). Maintenant on va conclure que (V.13) est vraie. Pour cela on écrit

$$\begin{aligned} & f(y) - f(x) - \langle Ax, y - x \rangle \\ &\leq -\langle Ay, x - y \rangle - \langle Ax, y - x \rangle \text{ avec (V.14) en échangeant les rôles de } x, y \\ &= \langle Ay - Ax, y - x \rangle \\ &\leq \|Ay - Ax\|\|y - x\| \text{ par Cauchy-Schwarz} \\ &\leq \|y - x\|^2 = o(\|y - x\|), \end{aligned}$$

où dans la dernière inégalité on a utilisé le fait que A est 1-Lipschitzienne. ■

V.III.3 Algorithme du gradient projeté

Ici on considère le problème de minimiser une fonction $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ sur une contrainte $C \subset \mathbb{R}^N$ convexe fermée non vide.

Au chapitre IV nous avons vu l'algorithme du gradient. Cet algorithme a la propriété que, à chaque itération, l'algorithme *progresse vers la solution*. On a vu que cela se traduit par :

$$f(x_{k+1}) \leq f(x_k)$$

Étant donné un point $x_k \in C$ dans la contrainte, on pourrait essayer de l'améliorer en y appliquant une étape de l'algorithme du gradient :

$$\hat{x}_{k+1} = x_k - \rho \nabla f(x_k).$$

En faisant cela, on obtient un point qui fait décroître la valeur de f . Mais rien ne garantit que \hat{x}_{k+1} soit encore dans C ! Or c'est un problème puisque on cherche le minimiseur de f sur C . On se retrouve donc avec un point \hat{x}_{k+1} sur les bras, qui est « bon » du point de vue de f , mais à priori mauvais vis-à-vis de C .

Une approche consiste alors à dire : au lieu de prendre \hat{x}_{k+1} , on va chercher parmi les points de C celui qui est le *plus proche* de \hat{x}_{k+1} , autrement dit la *projection* de \hat{x}_{k+1} sur C . Par définition il sera dans la contrainte, et comme il sera « pas trop loin » de \hat{x}_{k+1} , on espère qu'il aura la même propriété de faire décroître f (spoiler : oui).

Définition V.61 (Gradient projeté). Soit $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$, et $C \subset \mathbb{R}^N$ une contrainte convexe fermée non vide. L'**ALGORITHME DU GRADIENT PROJETÉ** appliqué à ce problème consiste à choisir un point initial $x_0 \in C$, puis à appliquer :

$$\begin{cases} \hat{x}_{k+1} = x_k - \rho \nabla f(x_k) \\ x_{k+1} = \text{proj}_C(\hat{x}_{k+1}) \end{cases}$$

En d'autres termes, l'algorithme du gradient projeté alterne une étape de l'algorithme du gradient par rapport à f , et une étape de projection par rapport à C . Comme son nom l'indique.

Remarque V.62 (Avantages et limitations de cette approche). L'efficacité de cette méthode est totalement dépendante de notre capacité à savoir projeter facilement, rapidement sur C . Il est illusoire de penser que ceci est possible pour tout ensemble, mais certaines contraintes comme le simplexe, l'orthant positif, des contraintes linéaires, peuvent être traitées en temps raisonnable.

Exemple V.63 (Contrainte linéaire sous forme standard). Considérons le problème de trouver un $x \in \mathbb{R}^N$ tel que

$$\Phi x = y \quad \text{et} \quad x \in \mathbb{R}_+^N,$$

où $\Phi \in \mathcal{M}_{M,N}(\mathbb{R})$ et $y \in \mathbb{R}^M$. Si ce problème admet une solution, alors il est équivalent à minimiser f sur C , où $C = \mathbb{R}_+^N$ et $f(x) := \frac{1}{2} \| \Phi x - y \|^2$. Dans ce cas l'algorithme du gradient projeté devient

$$x_{k+1} = (x_k - \rho \Phi^\top (\Phi x_k - y))_+, \quad \rho < \frac{2}{\|\Phi\|^2}.$$

Vérifions maintenant que cet algorithme est raisonnable, au sens où les solutions du problème sont des points stationnaires :

Proposition V.64 (Points fixes du gradient projeté). Soient $f \in \Gamma_0(\mathbb{R}^N)$ différentiable, $C \subset \mathbb{R}^N$ convexe fermé non vide, et $x^* \in \operatorname{argmin}_C f$. Alors $\operatorname{proj}_C(x^* - \rho \nabla f(x^*)) = x^*$ pour tout $\rho > 0$.

Démonstration. Au vu de la caractérisation de la projection par les angles (Proposition V.53), il nous suffit de montrer que

$$(\forall c \in C) \quad \langle x^* - \rho \nabla f(x^*) - x^*, c - x^* \rangle \leq 0.$$

Puisque $\rho > 0$, ceci est équivalent à montrer que

$$(\forall c \in C) \quad \langle \nabla f(x^*), c - x^* \rangle \geq 0.$$

Prenons donc un $c \in C$ quelconque. On peut alors calculer

$$\langle \nabla f(x^*), c - x^* \rangle = \lim_{t \rightarrow 0} \frac{f(x^* + t(c - x^*)) - f(x^*)}{t},$$

et cette fraction est bien positive ! En effet, pour $t \in]0, 1[$ on a par convexité que $x^* + t(c - x^*) \in C$, et puisque $x^* \in \operatorname{argmin}_C f$ on a forcément $f(x^* + t(c - x^*)) \geq f(x^*)$. D'où le résultat. ■

Vérifions que le gradient projeté fait bien décroître les valeurs de f :

Proposition V.65 (Décroissance de la méthode du gradient projeté). Soit $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$, pour $L > 0$, et $C \subset \mathbb{R}^N$ convexe fermé non vide. On considère la méthode du gradient projeté avec un pas constant $\rho \in]0, 2/L[$. Alors :

$$(\forall k \in \mathbb{N}) \quad f(x_{k+1}) \leq f(x_k).$$

Démonstration. On va commencer exactement comme pour la preuve de la Proposition IV.34, où l'on avait prouvé la chose suivante (IV.2) :

$$(\forall x, y \in \mathbb{R}^N) \quad f(y) - f(x) \leq \frac{L}{2} \|y - x\|^2 + \langle \nabla f(x), y - x \rangle. \quad (\text{V.15})$$

Avec $x = x_k$ et $y = x_{k+1}$, on a donc

$$f(x_{k+1}) - f(x_k) \leq \frac{L}{2} \|x_{k+1} - x_k\|^2 + \langle \nabla f(x_k), x_{k+1} - x_k \rangle.$$

En rappelant que $\hat{x}_{k+1} = x_k - \rho \nabla f(x_k)$, et en utilisant la caractérisation de la projection par les angles (Proposition V.53), on peut écrire

$$\begin{aligned} \|x_{k+1} - x_k\|^2 &= \langle \operatorname{proj}_C(\hat{x}_{k+1}) - x_k, \operatorname{proj}_C(\hat{x}_{k+1}) - x_k \rangle \\ &= \langle \operatorname{proj}_C(\hat{x}_{k+1}) - \hat{x}_{k+1}, \operatorname{proj}_C(\hat{x}_{k+1}) - x_k \rangle + \langle \hat{x}_{k+1} - x_k, \operatorname{proj}_C(\hat{x}_{k+1}) - x_k \rangle \\ &\leq \langle \hat{x}_{k+1} - x_k, \operatorname{proj}_C(\hat{x}_{k+1}) - x_k \rangle \quad (\text{Proposition V.53}) \text{ et } x_k \in C \\ &= -\rho \langle \nabla f(x_k), x_{k+1} - x_k \rangle. \end{aligned}$$

On a donc obtenu que

$$f(x_{k+1}) - f(x_k) \leq \left(1 - \frac{L\rho}{2}\right) \langle \nabla f(x_k), x_{k+1} - x_k \rangle.$$

D'une part, le fait que $\rho \in]0, 2/L[$ garantit que $1 - \frac{L\rho}{2} > 0$. D'autre part, la convexité de f et l'inégalité de l'hyperplan tangent, qui nous dit que $\langle \nabla f(x_k), x_{k+1} - x_k \rangle \leq f(x_{k+1}) - f(x_k)$, nous permet donc de conclure que

$$f(x_{k+1}) - f(x_k) \leq \left(1 - \frac{L\rho}{2}\right) (f(x_{k+1}) - f(x_k)),$$

d'où $f(x_{k+1}) - f(x_k) \leq 0$. ■

Nous énonçons maintenant le Théorème principal de cet algorithme :

Théorème V.66 (Convergence linéaire : Cas fortement convexe). Soit $f \in \Gamma_\mu(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$, pour $L \geq \mu > 0$, et $C \subset \mathbb{R}^N$ convexe fermé non vide. On note $x^* = \operatorname{argmin}_C f$, et on considère la méthode du gradient projeté avec un pas constant $\rho \in]0, 2/L[$. Alors $(\|x_k - x^*\|)_{k \in \mathbb{N}}$ converge linéairement, c'est-à-dire que :

$$(\exists \theta \in [0, 1]) (\forall x \in \mathbb{R}^N) \quad \|x_{k+1} - x^*\| \leq \theta \|x_k - x^*\|$$

Plus précisément, on peut montrer que

$$\theta = \max\{|1 - \rho\mu|; |1 - \rho L|\} = \begin{cases} 1 - \rho\mu & \text{si } \rho \leq \frac{2}{\mu+L} \\ \rho L - 1 & \text{si } \rho \geq \frac{2}{\mu+L}, \end{cases} \quad (\text{V.16})$$

qui est minimal lorsque $\rho = 2/(\mu + L)$. En particulier, x_k converge vers x^* .

La preuve de ce résultat va combiner deux ingrédients : les résultats sur l'algorithme du gradient vus au Chapitre IV, et la 1-Lipschitzianité de la projection :

Démonstration. On peut utiliser la propriété de Lipschitz de la projection (Théorème V.58), avec le fait que x^* est un point fixe de l'algorithme (Proposition V.64) :

$$\begin{aligned} \|x_{k+1} - x^*\| &= \|\operatorname{proj}_C(x_k - \rho \nabla f(x_k)) - \operatorname{proj}_C(x^* - \rho \nabla f(x^*))\| \\ &\leq \|(x_k - \rho \nabla f(x_k)) - (x^* - \rho \nabla f(x^*))\|. \end{aligned}$$

Or on a vu dans la preuve du Théorème IV.37 que l'étape de la méthode du gradient est θ -Lipschitzienne, ce qui permet de conclure que

$$\|x_{k+1} - x^*\| \leq \|(x_k - \rho \nabla f(x_k)) - (x^* - \rho \nabla f(x^*))\| \leq \theta \|x_k - x^*\|. \quad \blacksquare$$

Tout comme pour la méthode du gradient à pas constant, on a toujours convergence si f n'est que convexe. Dans ce cas on perd en vitesse de convergence, et on retombe sur une vitesse sous-linéaire (comparer avec le Théorème IV.44) :

Théorème V.67 (Convergence : cas convexe). Soient $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$, pour $L > 0$, et $C \subset \mathbb{R}^N$ convexe fermé non vide. On suppose que $\operatorname{argmin}_C f \neq \emptyset$. On considère la méthode du gradient projeté avec un pas constant $\rho \in]0, 2/L[$. Alors :

- 1) x_k converge vers $x^* \in \operatorname{argmin}_C f$,
- 2) $f(x_k) - \inf_C f = O\left(\frac{1}{k}\right)$.

Démonstration. Admis. Une preuve est disponible dans la Section A.II.3 de l'Annexe. ■

V.III.4 Algorithme de projection alternées *

Ici on s'intéresse au *problème de faisabilité*, qui consiste à être capable de trouver un point dans l'intersection de différentes contraintes :

$$\text{Trouver } x \in C := C_1 \cap \cdots \cap C_r. \quad (\text{V.17})$$

Pour ce genre de problèmes, typiquement chaque contrainte C_i est « simple », alors que C est plus compliquée.

Par exemple, on pourrait considérer que trouver la solution d'un système linéaire $Ax = b$ est difficile. Or cette égalité vectorielle est équivalente à vérifier des équations réelles (on note a_i les lignes de la matrice A) :

$$Ax = b \iff \forall i, \langle a_i, x \rangle = b_i.$$

Or, trouver une solution de $\langle a_i, x \rangle = b_i$ est très facile pour chaque i ! On sait même projeter sur cet hyperplan ! C'est trouver une solution *commune* qui est compliqué.

Un autre exemple consiste à dire que, ok, résoudre un système linéaire c'est facile, mais que pour des problèmes concrets on a souvent des contraintes naturelles qui s'ajoutent. Bien souvent, on veut que la solution de $Ax = b$ soit un vecteur de coordonnées positives. Autrement dit, on veut à la fois

$$Ax = b \quad \text{et} \quad x \in \mathbb{R}_+^N.$$

Pas facile à priori ! Faut-il/Peut-on modifier le pivot de Gauss pour garantir des coefficients positifs ? (non)

Donc dans cette section on va proposer un algorithme capable de résoudre le problème de faisabilité. L'idée est simple : on va projeter alternativement entre tous les C_i !

Définition V.68 (Algorithme de projection alternée). Soient $C_1, \dots, C_r \subset \mathbb{R}^N$ des ensembles convexes fermés non vides, et C leur intersection. On définit **L'ALGORITHME DE PROJECTION ALTERNÉE** ainsi :

$$\begin{cases} x_0 \in C_r, \\ x_{k+1} = (\text{proj}_{C_r} \circ \dots \circ \text{proj}_{C_1})(x_k). \end{cases}$$

Théorème V.69 (Convergence de la projection alternée). Soient C_1, \dots, C_r des ensembles convexes fermés non vides de \mathbb{R}^N , et $C = C_1 \cap \dots \cap C_r$ leur intersection. Si C est non vide, alors toute suite générée par l'algorithme de projection alternée converge vers un point de C .

Démonstration. Ici on notera p_C au lieu de proj_C pour simplifier. Soit $(x_k)_{k \in \mathbb{N}}$ la suite générée par l'algorithme de projection alternée, qui vérifie par définition $x_{k+1} = p_{C_r} \circ \dots \circ p_{C_1}(x_k)$. On va avoir besoin de donner un nom à toutes les suites intermédiaires, donc on définit pour tout $k \in \mathbb{N}$ et $i = 1, \dots, r$:

$$\hat{x}_k^0 := x_k \quad \text{et} \quad \hat{x}_k^i := p_{C_i}(\hat{x}_k^{i-1}).$$

Avec ces notations on voit que $\hat{x}_k^r = x_{k+1}$, et $\hat{x}_k^i \in C_i$.

Pour commencer, fixons un $c \in C$ quelconque. Puisque c est dans l'intersection, il vérifie pour tout i que $p_{C_i}(c) = c$. On peut donc utiliser le fait que la projection est non-expansive (Théorème V.58) pour écrire pour tout k :

$$\|x_{k+1} - c\| = \|\hat{x}_k^r - c\| = \|p_{C_r}(\hat{x}_k^{r-1}) - p_{C_r}(c)\| \leq \|\hat{x}_k^{r-1} - c\| \leq \dots \leq \|\hat{x}_k^0 - c\| = \|x_k - c\|.$$

On en déduit que, pour tout $i = 1, \dots, r$, la suite $(\|\hat{x}_k^i - c\|)_{k \in \mathbb{N}}$ est décroissante, et que toutes ces suites ont la même limite :

$$(\exists \ell \geq 0)(\forall i = 1, \dots, r) \quad \lim_{k \rightarrow +\infty} \|\hat{x}_k^i - c\| = \ell. \quad (\text{V.18})$$

On en déduit également que les suites $(\hat{x}_k^i)_{k \in \mathbb{N}}$ sont bornées, donc il existe une sous-suite k_n commune telle que toutes les sous-suites $(\hat{x}_{k_n}^i)_{k \in \mathbb{N}}$ soient convergentes. On notera x_∞^i leur limite, dont on sait que $x_\infty^i \in C_i$ puisque $x_k^i \in C_i$ et que les C_i sont fermés.

Maintenant, on utilise le Lemme V.57 avec $x = \hat{x}_k^i$ et $y = c$ pour obtenir :

$$\begin{aligned} \|x_k^{i+1} - x_k^i\|^2 &= \|(c - x_k^i) - (p_{C_i}(c) - p_{C_i}(x_k^i))\|^2 \\ &\leq \|c - x_k^i\|^2 - \|p_{C_i}(c) - p_{C_i}(x_k^i)\|^2 \\ &= \|c - x_k^i\|^2 - \|c - x_k^{i+1}\|^2. \end{aligned}$$

En passant à la limite lorsque $k_n \rightarrow +\infty$, on obtient alors

$$\|x_\infty^{i+1} - x_\infty^i\|^2 \leq \|c - x_\infty^i\|^2 - \|c - x_\infty^{i+1}\|^2 = \ell - \ell = 0,$$

ce qui veut dire que $x_\infty^{i+1} = x_\infty^i$. Ceci étant vrai pour tout i , on en déduit que toutes ces limites de sous-suites sont en fait le même point, que l'on note c_∞ , qui vérifie donc $c_\infty \in C$.

On peut maintenant conclure, en observant maintenant que si on prend $c = c_\infty$, alors $\lim_{k_n \rightarrow +\infty} \|\hat{x}_{k_n}^i - c_\infty\| = 0$. Autrement dit, la constante ℓ dans l'équation (V.18) est nulle. Ceci implique donc que c'est bien toute la suite x_k^i qui converge vers c_∞ . En particulier, $x_k = \hat{x}_{k-1}^r \rightarrow c_\infty$. ■

Remarque V.70 (Vitesse de convergence pour l'algorithme de projection alternée). Dans le Chapitre IV on s'est évertués à non seulement établir la convergence de la méthode du gradient mais aussi préciser quelle est sa vitesse de convergence, en fonction notamment de la difficulté du problème (à travers la valeur du conditionnement $\text{cond}(f)$).

Ici dans le Théorème V.69 nous n'avons que la convergence. Que peut-on alors dire des vitesses ? La réponse est hors-programme, mais voici quelques indications pour les plus curieuses. Essentiellement : l'algorithme converge plus vite lorsque le problème est « facile ».

- Un problème de faisabilité (V.17) est facile si les ensembles C_i « s'intersectent bien ». Plus précisément, lorsque ces ensembles ont une intersection **régulière**. Si on parle de contraintes d'égalités affines et d'inégalités convexes, alors « régulière » est à prendre au sens de la Définition V.33. Dans ce cas il est possible de montrer que les itérés x_k convergent vers une solution à vitesse linéaire, et que le taux de convergence linéaire θ dépend de l'angle formé entre les ensembles aux points où ils s'intersectent.
- De manière générale, lorsque l'intersection n'est pas régulière, la convergence des itérés peut être arbitrairement lente : sans hypothèse, on ne peut pas garantir de vitesse de convergence pour les itérés.

Remarque V.71 (La projection alternée est un gradient projeté). Considérons le problème de trouver un $x \in C \cap D \neq \emptyset$, où C et D sont deux ensembles convexes fermés. Alors ce problème est équivalent à minimiser f sur C , où $f(x) = \frac{1}{2} \text{dist}(x, D)^2$. D'après la Proposition V.60, on a pour cette fonction que $\text{Lip}(\nabla f) = 1$ et $\nabla f(x) = x - \text{proj}_D(x)$. Alors l'algorithme du gradient projeté s'écrit dans ce cas

$$x_{k+1} = \text{proj}_C(x_k - \rho(\nabla f(x_k))), \quad 0 < \rho < 2.$$

On observe qu'en prenant un pas court $\rho = 1$, on obtient $x_{k+1} = \text{proj}_C(\text{proj}_D(x_k))$ qui est exactement l'algorithme de la projection alternée pour une intersection de deux contraintes ! Qu'est-ce que cela implique du point de vue des vitesses de convergence ? Ici f n'est pas fortement convexe⁷ donc on doit appliquer le Théorème V.67. On obtient alors que les valeurs $f(x_k) = \text{dist}(x_k, D)$ tendent vers 0 avec une vitesse $O(\frac{1}{k})$, et que les itérés tendent vers une solution $x^* \in C \cap D$. Mais on ne peut rien dire de plus sur la vitesse de $\|x_k - x^*\|$, ou de $\text{dist}(x_k, C \cap D)$, conformément à la Remarque V.70.

⁷A moins que D soit réduit à un singleton. Saurez-vous voir pourquoi c'est évident ?

Remarque V.72 (La projection alternée n'est pas un gradient projeté). En présence de deux contraintes, on a vu dans la Remarque V.71 que la méthode de projection alternée est un cas particulier du gradient projeté. Malheureusement cela n'est pas vrai en général. En effet il est possible de montrer que la méthode de projection alternée pour $r \geq 3$ contraintes ne peut en aucun cas être écrite comme un cas particulier de la méthode du gradient projeté avec un choix intelligent de f .⁸

V.III.5 Pour aller plus loin *

On conclut ce chapitre avec quelques remarques. Ce sont essentiellement des remarques d'ouverture, pour votre culture, qui sont totalement hors-programme.

Remarque V.73 (Au delà du gradient projeté). Un des problèmes évidents de la méthode du gradient projeté est qu'il faut savoir ... projeter! Comme on l'a vu en début de chapitre, la projection est très facile à calculer pour certains ensembles : boules euclidiennes, l'orthant positif. Mais il n'existe pas de recette générale « miracle » pour projeter sur un ensemble quelconque. Voici quelques classes de problèmes que l'on rencontre typiquement en pratique :

- Les problèmes de programmation linéaire⁹, où f est affine, et la contrainte est définie par des égalités et inégalités affines. Ces problèmes apparaissent naturellement dans les sciences de la décision et de la planification. Un exemple célèbre est le problème du transport optimal¹⁰. Dans ce cas on pourra utiliser l'algorithme du *simplexe*¹¹ (1947).
- Les problèmes de programmation quadratique, où cette fois-ci f est quadratique¹² (les contraintes restent affines). Une famille d'algorithmes très efficaces (et même optimales en un certain sens) pour les résoudre sont les méthodes dites de *point intérieur*¹³ (1980-1999). Elles sont d'ailleurs si efficaces qu'elles permettent également de résoudre des problèmes beaucoup plus difficiles (programmation semi-définie).
- Les problèmes de programmation convexe, où f est convexe et les contraintes sont des inégalités convexes et égalités affines. Dans ce cas le problème est trop général, mais selon la structure du problème on peut toujours trouver un algorithme adapté. Citons par exemple la famille des méthodes dites d'éclatement (algorithmes du gradient proximal, de Douglas-Rachford, ...), très en vogue depuis les années 2000 pour résoudre les problèmes de traitement d'image (défloutage d'image, augmenter la résolution, diminuer le bruit, etc..). Ces méthodes sont en particulier très efficaces pour résoudre

⁸C'est un résultat qui date de 2012, dû à J.-B. Baillon, P.L. Combettes et R. Cominetti.

⁹[https://fr.wikipedia.org/wiki/Optimisation_linaire](https://fr.wikipedia.org/wiki/Optimisation_lin%C3%A9aire)

¹⁰<https://images.math.cnrs.fr/Le-transport-numrique-et-ses-applications-Partie-1.html?lang=fr>

¹¹https://fr.wikipedia.org/wiki/Algorithme_du_simplexe

¹²https://fr.wikipedia.org/wiki/Optimisation_quadratique

¹³https://en.wikipedia.org/wiki/Interior-point_method

des problèmes non-lisses, comme par exemple le problème de régression parcimonieuse¹⁴ qui apparaît en traitement du signal, traitement de l'image, ainsi qu'en statistiques :

$$\min_{x \in \mathbb{R}^N} \alpha \|x\|_1 + \frac{1}{2} \|Ax - y\|_2^2.$$

Remarque V.74 (Problème général). Ici on vient de voir que la méthode du gradient projeté converge si f est lisse, convexe, et C convexe. Que se passe-t-il si ces hypothèses ne sont pas vérifiées ?

- f non convexe : Sans convexité, et même lorsqu'il n'y a pas de contrainte, cela se complique. Déjà on sait que même si on converge on risque d'être coincé dans un minimiseur local voire un point critique (cf. x^3). On sait également depuis longtemps (1950 environ) que sans convexité il est possible que l'algorithme du gradient ne converge pas : les trajectoires peuvent tourner en rond.¹⁵ Mais récemment (2005-2015) on s'est rendu compte que ce phénomène n'arrivait pas « souvent ». Pour des fonctions non-convexes « normales¹⁶ » (polynomiales par exemple) la convergence vers un point critique est garantie.
- C non convexe : Dans ce cas, la projection n'est plus définie de manière unique (cf. Figure V.19). Mais on pourrait toujours implémenter l'algorithme en prenant à chaque itération « une projection » quelconque. Dans ce cas on a les mêmes résultats que pour f non convexe : la convergence vers un point critique du problème est garantie pour des ensembles « normaux ».

¹⁴[https://fr.wikipedia.org/wiki/Lasso_\(statistiques\)](https://fr.wikipedia.org/wiki/Lasso_(statistiques))

¹⁵Les plus curieux pourront aller regarder ce GIF qui illustre ce fait avec une fonction non-convexe connue sur le nom de « mexican hat » : https://raw.githubusercontent.com/Guillaume-Garrigos/guillaume-garrigos.github.io/master/assets/math/images/mex_trajectoire.gif

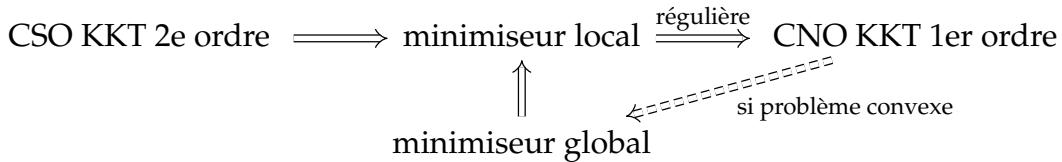
¹⁶La définition de « normale » est un peu compliquée, mais pourrait être résumée par : « sa définition ne contient rien qui ait à voir de près ou de loin avec $\sin(x)$. »

V.IV Récapitulatif du Chapitre V

On considère une fonction $f : \mathbb{R}^N \rightarrow \mathbb{R}$, et un ensemble de contraintes défini par des inégalités et égalités :

$$C = \{x \in \mathbb{R}^N \mid g_1(x) \leq 0, \dots, g_p(x) \leq 0, h_1(x) = 0, \dots, h_q(x)\}$$

et le problème d'optimisation associé : $\text{minimiser}_{x \in C} f(x)$. Les Théorèmes de Lagrange-KKT nous donnent les implications suivantes :



Condition Nécessaire d'Optimalité de KKT du 1er ordre

$$\left\{ \begin{array}{ll} \nabla f(x) + \sum_{i=1}^p \alpha_i \nabla g_i(\bar{x}) + \sum_{j=1}^q \beta_j \nabla h_j(\bar{x}) = 0 & (\text{Condition de stationnarité}) \\ \forall i = 1, \dots, p \quad g_i(\bar{x}) \leq 0 & (\text{Condition d'admissibilité : inégalités}) \\ \forall j = 1, \dots, q \quad h_j(\bar{x}) = 0 & (\text{Condition d'admissibilité : égalités}) \\ \forall i = 1, \dots, p \quad \alpha_i \geq 0 & (\text{Multiplicateur : inégalités}) \\ \forall i = 1, \dots, p \quad \alpha_i g_i(\bar{x}) = 0 & (\text{Condition de complémentarité : inégalités}) \end{array} \right.$$

Pour que l'implication « minimiseur local \implies CNO KKT 1er ordre » ait lieu, il faut que la contrainte soit **régulière** en \bar{x} , c'est-à-dire qu'elle vérifie l'une des deux propriétés :

- **linéaire** (les g_i et h_j sont affines);
- **qualifiée** : $\{\nabla g_i(x), \nabla h_j(x)\}_{\substack{i \in I(\bar{x)} \\ 1 \leq j \leq q}}$ est libre. $I(\bar{x})$ désigne les contraintes **actives** en \bar{x} .

Condition Suffisante d'Optimalité de KKT du 2e ordre

- 1) La CNO de KKT du 1er ordre est vérifiée, avec des multiplicateurs α_i, β_j ;
- 2) La Hessienne Lagrangienne $\nabla^2 f(\bar{x}) + \sum_{i=1}^p \alpha_i \nabla^2 g_i(\bar{x}) + \sum_{j=1}^q \beta_j \nabla^2 h_j(\bar{x})$ est $\succ 0$;
- 3) La complémentarité stricte : $\alpha_i \neq 0 \Leftrightarrow g_i(\bar{x}) = 0$ pour $i = 1, \dots, p$.

Annexe A

Convexité(s) et Convergence de méthodes de descente

Sommaire

A.I	Un peu plus d'Analyse variationnelle	124
A.I.1	Convexité(s) et monotonie(s)	124
A.I.2	Caractérisation de la convexité via la Hessienne	126
A.I.3	Lipschitzianité et cocoercivité	127
A.II	Convergence(s) de la méthode du gradient	130
A.II.1	Méthode du gradient : cas fortement convexe non C^2	130
A.II.2	Méthode du gradient : cas convexe	132
A.II.3	Méthode du gradient projeté : cas convexe	137
A.II.4	Méthode du gradient optimal	139

Dans cette annexe nous commençons par montrer quelques caractérisations supplémentaires de la convexité, forte convexité, et Lipschitzianité du gradient. Cela nous permet dans un second temps de prouver des résultats laissés admis jusque là, ou tout simplement de donner une preuve plus directe à certains Théorèmes :

- Preuve directe de la caractérisation de la convexité via la Hessienne qui ne nécessite pas de passer par le cas univarié, comme cela est fait dans la preuve du Théorème III.33. Cf. Section A.I.2.
- Preuve de la convergence linéaire des *itérés* pour la méthode du gradient, pour les fonctions fortement convexes (Théorème IV.37), sans faire l'hypothèse que la fonction f est de classe C^2 . Cf. Section A.II.1.
- Preuve de la convergence linéaire des *valeurs* pour la méthode du gradient, pour les fonctions fortement convexes (Théorème IV.42). Cf. Section A.II.1.

- Preuves de la convergence sous-linéaire de la méthode du gradient (projeté) pour les fonctions convexes (Théorèmes IV.44 et V.67). Cf. Sections A.II.2 et A.II.3.
- Preuve de la convergence linéaire de la méthode du gradient à pas optimal, pour les fonctions fortement convexes (Théorème IV.57). Cf. Section A.II.4.

A.I Un peu plus d'Analyse variationnelle

A.I.1 Convexité(s) et monotonie(s)

Remarque A.1 (Croissance et Monotonie). Pour les fonctions univariées, on a vu dans la Proposition III.13 que la convexité était équivalent à la croissance de la dérivée. Or il n'y a pas de notion de « croissance » pour le gradient, car la relation d'ordre canonique sur \mathbb{R}^N n'est pas un ordre total. Mais il existe une notion un peu plus générale, celle de fonction **monotone**. En effet, la croissance d'une fonction univariée $f : \mathbb{R} \rightarrow \mathbb{R}$ s'écrit

$$(\forall x, y \in \mathbb{R}) \quad x \leq y \Rightarrow f(x) \leq f(y).$$

Cette propriété est en fait équivalente à dire que $y - x$ et $f(y) - f(x)$ ont le même signe. Autrement dit :

$$(\forall x, y \in \mathbb{R}) \quad (f(y) - f(x))(y - x) \geq 0.$$

On peut alors étendre cette relation aux champs de vecteurs, et dire que $F : \mathbb{R}^N \rightarrow \mathbb{R}^N$ est monotone si :

$$(\forall x, y \in \mathbb{R}) \quad \langle F(y) - F(x), y - x \rangle \geq 0.$$

On peut alors montrer (Proposition suivante) que la convexité d'une fonction $f : \mathbb{R}^N \rightarrow \mathbb{R}$ est équivalente à la monotonie de son gradient $\nabla f : \mathbb{R}^N \rightarrow \mathbb{R}^N$.

Proposition A.2 (Convexité via le gradient). Soit $f : U \subset \mathbb{R}^N \rightarrow \mathbb{R}$ une fonction différentiable sur U , et $C \subset U$ convexe non vide. Les propriétés suivantes sont alors équivalentes :

- i) f est convexe sur C , c'est à dire $f \in \Gamma_0(C)$;
- ii) $(\forall (x, y) \in C^2) \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$;
- iii) $(\forall (x, y) \in C^2) \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0$.

Démonstration. i) \Rightarrow ii). Soient $(x, y) \in C^2$ quelconques. Pour $\alpha \in]0, 1[$, on pose $z_\alpha := (1 - \alpha)x + \alpha y$. On a alors $f(z_\alpha) \leq (1 - \alpha)f(x) + \alpha f(y) = f(x) + \alpha(f(y) - f(x))$, donc

$$f(y) - f(x) \geq \frac{1}{\alpha}(f(z_\alpha) - f(x)) \xrightarrow{\alpha \rightarrow 0^+} Df(x)(y - x) = \langle \nabla f(x), y - x \rangle.$$

ii) \Rightarrow i) : On a

$$f(x) \geq f(z_\alpha) + \langle \nabla f(z_\alpha), x - z_\alpha \rangle \tag{A.1}$$

$$f(y) \geq f(z_\alpha) + \langle \nabla f(z_\alpha), y - z_\alpha \rangle. \tag{A.2}$$

En sommant $(1 - \alpha)$ fois la relation (A.1) et α fois la relation (A.2), et en utilisant le fait que $(1 - \alpha)(x - z_\alpha) + \alpha(y - z_\alpha) = 0$, on obtient l'inégalité de convexité.

ii) \Rightarrow iii) : On écrit

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle \\ f(x) &\geq f(y) + \langle \nabla f(y), x - y \rangle. \end{aligned}$$

En sommant on obtient l'inégalité désirée.

iii) \Rightarrow ii) : Soit $g(t) := f((1-t)x + ty)$ pour $t \in [0, 1]$. On remarque que $g'(t) = \langle \nabla f(z_t), y - x \rangle$, et en particulier que $g'(0) = \langle \nabla f(x), y - x \rangle$. Donc il nous suffit de montrer que $g(1) - g(0) - g'(0) \geq 0$. D'après notre hypothèse, on a

$$g'(t) - g'(0) = \langle \nabla f(z_t) - \nabla f(x), y - x \rangle = \frac{1}{t} \langle \nabla f(z_t) - \nabla f(x), z_t - x \rangle \geq 0.$$

D'autre part, comme g est continue sur $[0, 1]$ et dérivable sur $]0, 1[$, on peut utiliser le théorème des accroissements finis qui nous dit qu'il existe $c \in]0, 1[$ tel que $\frac{g(1) - g(0)}{1} = g'(c)$. En combinant ces deux résultats, on en déduit que $g(1) - g(0) \geq g'(0)$, ce qui donne l'inégalité désirée. ■

Un analogue à la Proposition A.2 pour les fonctions fortement convexes :

Proposition A.3 (Forte convexité via gradient). Soient $C \subset \mathbb{R}^N$ convexe, $f : C \rightarrow \mathbb{R}$ une fonction différentiable en tout point de C , et $\mu > 0$. Les propriétés suivantes sont alors équivalentes :

- i) f est fortement convexe sur C , c'est à dire $f \in \Gamma_\mu(C)$;
- ii) $(\forall (x, y) \in C^2) \quad f(y) - f(x) - \langle \nabla f(x), y - x \rangle \geq \frac{\mu}{2} \|y - x\|^2$;
- iii) $(\forall (x, y) \in C^2) \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \mu \|y - x\|^2$.

Démonstration. Soit $\mu > 0$ et $f = g + (\mu/2)\|\cdot\|^2$. En particulier on a $\nabla f(x) = \nabla g(x) + \mu x$ sur C .

i) \Leftrightarrow ii) On peut écrire :

$$\begin{aligned} &f(y) - f(x) - \langle \nabla f(x), y - x \rangle \\ &= g(y) - g(x) - \langle \nabla g(x), y - x \rangle + (\mu/2)\|y\|^2 - (\mu/2)\|x\|^2 - \langle \mu x, y - x \rangle \\ &= g(y) - g(x) - \langle \nabla g(x), y - x \rangle + (\mu/2)\|y - x\|^2. \end{aligned}$$

On conclut donc avec les Propositions III.30 et A.2.

i) \Leftrightarrow iii) On peut écrire :

$$\begin{aligned} &\langle \nabla f(y) - \nabla f(x), y - x \rangle \\ &= \langle \nabla g(y) - \nabla g(x), y - x \rangle + \mu \|y - x\|^2. \end{aligned}$$

On conclut donc avec les Propositions III.30 et A.2. ■

Un résultat en quelque sorte « dual » de la Proposition A.3.

Proposition A.4 (Forte convexité via gradient II). Soit $f \in \Gamma_\mu(\mathbb{R}^N)$ une fonction différentiable, avec $\mu > 0$. Alors les propriétés suivantes ont lieu :

- 1) $(\forall x, y \in \mathbb{R}^N) \quad \frac{1}{2\mu} \|\nabla f(y) - \nabla f(x)\|^2 \geq f(y) - f(x) - \langle \nabla f(x), y - x \rangle;$
- 2) $(\forall x, y \in \mathbb{R}^N) \quad \frac{1}{\mu} \|\nabla f(y) - \nabla f(x)\|^2 \geq \langle \nabla f(y) - \nabla f(x), y - x \rangle.$

Démonstration. i) (voir [15, Theorem 2.1.10]) Soit $x \in \mathbb{R}^N$ fixé, et soit $\phi(y) := f(y) - \langle \nabla f(x), y \rangle$. Puisque $f \in \Gamma_\mu(\mathbb{R}^N)$ alors $\phi \in \Gamma_\mu(\mathbb{R}^N)$ aussi, comme somme d'un fonction fortement convexe et d'une fonction convexe (car linéaire). On calcule $\nabla \phi(y) = \nabla f(y) - \nabla f(x)$, et on en déduit que $\text{argmin} \phi = \{x\}$. On peut donc écrire d'après ii) que pour tout $y \in \mathbb{R}^N$:

$$\phi(x) = \min_{v \in \mathbb{R}^N} \phi(v) \geq \min_{v \in \mathbb{R}^N} \phi(y) + \langle \nabla \phi(y), v - y \rangle + \frac{\mu}{2} \|v - y\|^2.$$

Or le terme de droite est un problème d'optimisation en v , fortement convexe, dont l'unique solution v^* vérifie la CNO du 1er ordre : $\nabla \phi(y) + \mu(v^* - y) = 0$. Autrement dit, $v^* = y - \frac{1}{\mu} \nabla \phi(y)$. On a donc

$$\begin{aligned} \phi(x) &\geq \phi(y) + \langle \nabla \phi(y), v^* - y \rangle + \frac{\mu}{2} \|v^* - y\|^2 \\ &= \phi(y) - \frac{1}{\mu} \|\nabla \phi(y)\|^2 + \frac{1}{2\mu} \|\nabla \phi(y)\|^2 \\ &= \phi(y) - \frac{1}{2\mu} \|\nabla \phi(y)\|^2 \end{aligned}$$

On a donc bien montré que

$$\frac{1}{2\mu} \|\nabla \phi(y)\|^2 \geq \phi(y) - \phi(x),$$

où $\phi(y) - \phi(x) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle$ et $\nabla \phi(y) = \nabla f(y) - \nabla f(x)$.

ii) Il suffit d'appliquer i), puis de nouveau i) en inversant les rôles de x et y , puis d'en faire la somme. ■

A.I.2 Caractérisation de la convexité via la Hessienne

Rappelons ici le Théorème III.19 qui caractérise la convexité avec la positivité de la Hessienne :

Théorème A.5 (Convexité via Hessienne). Soit $f : U \subset \mathbb{R}^N \rightarrow \mathbb{R}$, deux fois différentiable sur U , et $C \subset U$ convexe et ouvert. Alors les propriétés suivantes sont équivalentes :

- i) f est convexe sur C , c'est à dire $f \in \Gamma_0(C)$;

$$ii) (\forall x \in C) \quad \nabla^2 f(x) \succeq 0.$$

Voici une preuve directe de ce résultat, qui ne passe pas par le cas univarié étudié dans la Section III.I.3, mais utilise plutôt la monotonie du gradient :

Démonstration.

ii) \Rightarrow i) : Soit $x \in C$, et $d \in \mathbb{R}^N$ quelconque ; il nous faut montrer que $\langle \nabla^2 f(x)d, d \rangle \geq 0$. D'après la Proposition I.78.iii), on a $\nabla^2 f(x) = J(\nabla f)(x)$, donc :

$$\langle \nabla^2 f(x)d, d \rangle = \langle J(\nabla f)(x)d, d \rangle = d^T J(\nabla f)(x)d.$$

D'autre part, $J(\nabla f)(x)d$ est la dérivée directionnelle de ∇f en x dans la direction d , donc :

$$\begin{aligned} d^T J(\nabla f)(x)d &= d^T \lim_{t \rightarrow 0} \frac{\nabla f(x + td) - \nabla f(x)}{t} = \lim_{t \rightarrow 0} \frac{\langle d, \nabla f(x + td) - \nabla f(x) \rangle}{t} \\ &= \lim_{t \rightarrow 0} \frac{\langle (x + td) - x, \nabla f(x + td) - \nabla f(x) \rangle}{t^2} \geq 0, \end{aligned}$$

la dernière inégalité provenant de la Proposition A.2.iii), et du fait que pour t suffisamment petit, on a $x + td \in C$ puisque C est ouvert.

i) \Rightarrow ii) : Soient $x, y \in C$ fixés. Soit $g : U \rightarrow \mathbb{R}$ définie par $g(z) = \langle \nabla f(z), y - x \rangle$. Elle est différentiable comme f , et $\nabla g(z) = \nabla^2 f(z)(y - x)$. En utilisant le Théorème de Taylor-Lagrange, on sait qu'il existe $z_\alpha \in]x, y[$ tel que :

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle = g(y) - g(x) = \langle \nabla g(z_\alpha), y - x \rangle = \langle \nabla^2 f(z_\alpha)(y - x), y - x \rangle \geq 0,$$

où $\alpha \in]0, 1[$, et la dernière inégalité vient de l'hypothèse, et du fait que $z_\alpha \in C$ par convexité. On conclut donc avec la Proposition A.2. ■

A.I.3 Lipschitzianité et cocoercivité

Quelques caractérisations de $C_L^{1,1}(\mathbb{R}^N)$ pour les fonctions convexes, qui ne font pas intervenir l'hypothèse de double différentiabilité :

Proposition A.6 (Lipschitzianité via le gradient). *Soit $f \in \Gamma_0(\mathbb{R}^N)$ différentiable. Alors les propriétés suivantes sont équivalentes :*

- i) ∇f est L -Lipschitzien : $(\forall x, y \in \mathbb{R}^N) \quad \|\nabla f(y) - \nabla f(x)\| \leq L\|y - x\|$.
- ii) $(\forall x, y \in \mathbb{R}^N) \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle \leq L\|y - x\|^2$.
- iii) $(\forall x, y \in \mathbb{R}^N) \quad f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|y - x\|^2$.
- iv) $(\forall x, y \in \mathbb{R}^N) \quad \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle$.
- v) $(\forall x, y \in \mathbb{R}^N) \quad \frac{1}{L}\|\nabla f(y) - \nabla f(x)\|^2 \leq \langle \nabla f(y) - \nabla f(x), y - x \rangle$.

Remarque A.7 (Cocoercivité). La propriété v) est bien plus forte et précise que la simple monotonie de ∇f (voir Proposition A.2.iii)). Cette propriété s'appelle la **cocoercivité** de ∇f . Plus précisément, on dit que ∇f est $\frac{1}{L}$ -cocoercive. L'équivalence entre « ∇f est Lipschitzienne » et « ∇f est cocoercive » est connue sous le nom du Théorème de Baillon-Haddad [16, Theorem 3.13].

Remarque A.8 (Dualité). Si on compare la Proposition A.6 avec les Propositions A.3 et A.4, on voit qu'il y a beaucoup de propriétés similaires, mais en fait opposées. Par exemple les termes en $\nabla f(y) - \nabla f(x)$ s'échangent avec des termes en $\|y - x\|$ et μ s'échange avec $\frac{1}{L}$. C'est en fait une conséquence d'un principe de *dualité* entre $\Gamma_\mu(\mathbb{R}^N)$ et $C_L^{1,1}(\mathbb{R}^N)$, qui n'est pas au programme.

Démonstration. (Voir [16, Lemma 1.30] ou [15, Theorem 2.1.5])

i) \Rightarrow ii) : Il suffit d'utiliser l'inégalité de Cauchy-Schwarz, et i).

ii) \Rightarrow iii) : Soit $x, y \in \mathbb{R}^N$, et posons $g(t) = f(z_t)$ où $z_t = (1-t)x + ty$. Alors $g'(t) = \langle \nabla f(z_t), y - x \rangle$, et :

$$\begin{aligned} f(y) - f(x) - \langle \nabla f(x), y - x \rangle &= g(1) - g(0) - \langle \nabla f(x), y - x \rangle \\ &= \int_0^1 g'(t) dt - \langle \nabla f(x), y - x \rangle \\ &= \int_0^1 \langle \nabla f(z_t) - \nabla f(x), y - x \rangle dt \\ &\leq \int_0^1 Lt \|y - x\|^2 = \frac{L}{2} \|y - x\|^2. \end{aligned}$$

iii) \Rightarrow iv) : Soit $x, y \in \mathbb{R}^N$, et posons $g(y) = f(y) - \langle \nabla f(x), y \rangle$. Puisque $\nabla g(y) = \nabla f(y) - \nabla f(x)$, on en déduit que $g \in C_L^{1,1}(\mathbb{R}^N)$. De plus, g est la somme d'une fonction convexe et d'une forme linéaire, donc elle est convexe aussi. On voit que $\nabla g(x) = 0$, donc $x \in \operatorname{argmin} g$. On peut appliquer maintenant iii) à g , en les points $y - \frac{1}{L}\nabla g(y)$ et y :

$$\begin{aligned} g(y - \frac{1}{L}\nabla g(y)) - g(y) - \langle \nabla g(y), -\frac{1}{L}\nabla g(y) \rangle &\leq \frac{L}{2} \left\| -\frac{1}{L}\nabla g(y) \right\|^2 \\ \Leftrightarrow g(y - \frac{1}{L}\nabla g(y)) - g(y) + \frac{1}{L} \|\nabla f(y) - \nabla f(x)\|^2 &\leq \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2 \end{aligned}$$

Puisque $x \in \operatorname{argmin} g$, donc on obtient :

$$\begin{aligned} \Rightarrow g(x) - g(y) + \frac{1}{L} \|\nabla f(y) - \nabla f(x)\|^2 &\leq \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2 \\ \Leftrightarrow f(x) - f(y) - \langle \nabla f(x), x - y \rangle &\leq \frac{-1}{2L} \|\nabla f(y) - \nabla f(x)\|^2. \end{aligned}$$

iv) \Rightarrow v) : Il suffit d'utiliser iv) deux fois d'affilée en inversant les rôles de x et y , et de faire la somme.

v) \Rightarrow i) : Il suffit d'utiliser l'inégalité de Cauchy-Schwarz, et de diviser par $\|\nabla f(y) - \nabla f(x)\|$. ■

Proposition A.9. Soit $f \in \Gamma_\mu(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ avec $\mu, L > 0$. Alors $L \geq \mu$.

Démonstration. D'après les Propositions A.3.ii) et A.6.iii), on a

$$(\forall x, y \in \mathbb{R}^N) \quad \frac{\mu}{2} \|y - x\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2} \|y - x\|^2. \quad (\text{A.3})$$
■

Proposition A.10. Soit $f \in \Gamma_\mu(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ avec $\mu = L$. Alors il existe $b \in \mathbb{R}^N, c \in \mathbb{R}$ tels que $f(x) = \frac{\mu}{2} \|x\|^2 + \langle b, x \rangle + c$.

Démonstration. On reprend (A.3) où les inégalités deviennent ici des égalités, et on conclut avec $b = \nabla f(0)$ et $c = f(0)$. ■

Remarque A.11 (Γ_μ et $C_L^{1,1}$ combinés). Lorsque on a une fonction dans $\Gamma_\mu(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$, on peut combiner leurs propriétés ! Par exemple en combinant Proposition A.6.v) et A.6.iii), on obtient

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{\mu}{2} \|y - x\|^2 + \frac{1}{2L} \|\nabla f(y) - \nabla f(x)\|^2.$$

Mais le fait est que l'on a ici en quelque sorte utilisé séparément la forte convexité et le gradient Lipschitz. Lorsque les deux sont réunis, on peut obtenir des constantes un peu meilleures (ce qui aura de l'importance par la suite).

Proposition A.12. Soit $f \in \Gamma_\mu(\mathbb{R}^N)$ avec $\mu > 0$, et soit $L > \mu$. Alors les propriétés suivantes sont équivalentes :

- i) ∇f est L -Lipschitzien.
- ii) $(\forall x, y \in \mathbb{R}^N) \quad \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{\mu L}{\mu + L} \|y - x\|^2 + \frac{1}{\mu + L} \|\nabla f(y) - \nabla f(x)\|^2$.

Démonstration. (voir [15, Theorem 2.1.12]) Soit $g(x) = f(x) - \frac{\mu}{2} \|x\|^2$. Puisque $f \in \Gamma_\mu(\mathbb{R}^N)$ alors $g \in \Gamma_0(\mathbb{R}^N)$ d'après la Proposition III.30. On peut également écrire que $f \in C_L^{1,1}(\mathbb{R}^N) \Leftrightarrow g \in C_{L-\mu}^{1,1}(\mathbb{R}^N)$, ceci découlant de

$$0 \leq g(y) - g(x) - \langle \nabla g(x), y - x \rangle = f(y) - f(x) - \langle \nabla f(x), y - x \rangle - \frac{\mu}{2} \|y - x\|^2$$

et de la Proposition A.6. On considère maintenant deux cas :

Cas $\mu = L$: Dans ce cas on a i) $\Leftrightarrow g \in C_0^{1,1}(\mathbb{R}^N)$, ce qui est équivalent à dire que ∇g est constante. D'autre part, ii) est équivalente à :

$$\begin{aligned} & \langle \frac{1}{\sqrt{\mu}}(\nabla f(y) - \nabla f(x)), \sqrt{\mu}(y - x) \rangle \geq \frac{1}{2} \|\sqrt{\mu}(y - x)\|^2 + \frac{1}{2} \|\frac{1}{\sqrt{\mu}}(\nabla f(y) - \nabla f(x))\|^2 \\ \Leftrightarrow & 0 \geq \frac{1}{2} \|\sqrt{\mu}(y - x) - \frac{1}{\sqrt{\mu}}(\nabla f(y) - \nabla f(x))\|^2 \\ \Leftrightarrow & \sqrt{\mu}(y - x) = \frac{1}{\sqrt{\mu}}(\nabla f(y) - \nabla f(x)) \\ \Leftrightarrow & \nabla g(y) = \nabla g(x), \end{aligned}$$

cette dernière propriété voulant dire que ∇g est constante.

Cas $L > \mu$: On utilise la Proposition A.6.v) pour écrire que $g \in C_{L-\mu}^{1,1}(\mathbb{R}^N)$ est équivalent à, pour tout $x, y \in \mathbb{R}^N$:

$$\begin{aligned} & \langle \nabla g(y) - \nabla g(x), y - x \rangle \geq \frac{1}{L-\mu} \|\nabla g(y) - \nabla g(x)\|^2 \\ \Leftrightarrow & \langle \nabla f(y) - \nabla f(x), y - x \rangle - \mu \|y - x\|^2 \geq \frac{1}{L-\mu} \|\nabla f(y) - \nabla f(x) - \mu(y - x)\|^2 \\ \Leftrightarrow & (1 + \frac{2\mu}{L-\mu}) \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{L-\mu} \|\nabla f(y) - \nabla f(x)\|^2 + (\mu + \frac{\mu^2}{L-\mu}) \|y - x\|^2 \\ \Leftrightarrow & \frac{L+\mu}{L-\mu} \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \frac{1}{L-\mu} \|\nabla f(y) - \nabla f(x)\|^2 + \frac{\mu L}{L-\mu} \|y - x\|^2, \end{aligned}$$

qui est équivalent à ii). ■

A.II Convergence(s) de la méthode du gradient

A.II.1 Méthode du gradient : cas fortement convexe non C^2

Ici on prouve le Théorème IV.37 sur la convergence linéaire de la méthode du gradient pour une fonction fortement convexe, sans utiliser l'hypothèse que f est de classe C^2 .

Démonstration du Théorème IV.37. (Voir [15, Theorem 2.1.15] ou [17, Theorem 3.1]) Soit $x \in \mathbb{R}^N$ et $x^+ = x - \rho \nabla f(x)$. On utilise le fait que $\nabla f(x^*) = 0$ (voir Théorème II.9) pour écrire :

$$\|x^+ - x^*\|^2 = \|x - x^* - \rho \nabla f(x)\|^2 = \|x - x^*\|^2 + \rho^2 \|\nabla f(x)\|^2 - 2\rho \langle x - x^*, \nabla f(x) \rangle.$$

Comme $\rho \in]0, 2/L[$, on peut écrire $\rho = 2\alpha/(\mu + L)$, où $\alpha \in]0, 1 + \mu/L[$. On écrit alors :

$$\|x^+ - x^*\|^2 = \|x - x^*\|^2 + \rho^2 \|\nabla f(x)\|^2 - 2\rho\alpha \langle x - x^*, \nabla f(x) \rangle - 2\rho(1 - \alpha) \langle x - x^*, \nabla f(x) \rangle. \quad (\text{A.4})$$

Puisque f est fortement convexe et à gradient Lipschitzien, on peut utiliser la caractérisation de la Proposition A.12.ii) (on rappelle que $\nabla f(x^*) = 0$) :

$$\langle \nabla f(x), x - x^* \rangle \geq \frac{\mu L}{\mu + L} \|x - x^*\|^2 + \frac{1}{\mu + L} \|\nabla f(x)\|^2.$$

En insérant cette inégalité dans (A.4) (sur le terme proportionnel à $\alpha > 0$), et en utilisant la définition de α , on obtient :

$$\begin{aligned} \|x^+ - x^*\|^2 &\leq \|x - x^*\|^2 \left(1 - \alpha 2\rho \frac{\mu L}{\mu + L}\right) + \|\nabla f(x)\|^2 \left(\rho^2 - \alpha 2\rho \frac{1}{\mu + L}\right) \\ &\quad - 2\rho(1 - \alpha) \langle x - x^*, \nabla f(x) \rangle \\ &= \|x - x^*\|^2 (1 - \mu L \rho^2) - 2\rho(1 - \alpha) \langle x - x^*, \nabla f(x) \rangle. \end{aligned}$$

On va maintenant majorer le dernier terme proportionnel à $(1 - \alpha)$, dont on ne connaît pas le signe. Puisque f est fortement convexe et à gradient Lipschitzien, on peut utiliser les Propositions A.3.iii) et A.6.ii) pour écrire

$$(\forall x \in \mathbb{R}^N) \quad \mu \|x - x^*\|^2 \leq \langle \nabla f(x), x - x^* \rangle \leq L \|x - x^*\|^2. \quad (\text{A.5})$$

On considère maintenant deux cas :

Cas $\rho \leq 2/(\mu + L)$: Ici on a $(1 - \alpha) \geq 0$. On peut donc utiliser la première inégalité de (A.5) pour écrire

$$\|x^+ - x^*\|^2 \leq \|x - x^*\|^2 (1 - \mu L \rho^2) - 2\rho(1 - \alpha) \langle x - x^*, \nabla f(x) \rangle \leq \theta^2 \|x - x^*\|^2,$$

où $\theta^2 = 1 - \mu L \rho^2 - 2\rho(1 - \alpha)\mu = (1 - \rho\mu)^2$.

Cas $\rho \geq 2/(\mu + L)$: Ici on a $(1 - \alpha) \leq 0$. On peut donc utiliser la deuxième inégalité de (A.5) pour écrire

$$\|x^+ - x^*\|^2 \leq \|x - x^*\|^2 (1 - \mu L \rho^2) - 2\rho(1 - \alpha) \langle x - x^*, \nabla f(x) \rangle \leq \theta^2 \|x - x^*\|^2,$$

où $\theta^2 = 1 - \mu L \rho^2 - 2\rho(1 - \alpha)L = (1 - \rho L)^2$. ■

On passe ensuite Au Théorème IV.42, qui porte sur la vitesse de convergence linéaire de $f(x_k) - \inf f$. Nous proposons ici une preuve simplifiée : nous allons montrer que $f(x_k) - \inf f$ converge linéairement, mais nous n'allons pas vérifier que le taux de contraction est exactement le même θ que celui du Théorème IV.37.iii). On aura un θ un peu moins bon.

Démonstration du Théorème IV.42 avec un θ quelconque. Reprenons la preuve de la Proposition IV.34, où l'on avait montré que :

$$f(x^+) - \inf f \leq f(x) - \inf f - \left(\rho - \frac{L\rho^2}{2}\right) \|\nabla f(x)\|^2.$$

Puisque f est fortement convexe, on peut utiliser la Proposition A.4.i) qui nous donne :

$$f(x) - \inf f = f(x) - f(x^*) - \langle \nabla f(x^*), x - x^* \rangle \leq \frac{1}{2\mu} \|\nabla f(x) - \nabla f(x^*)\|^2 = \frac{1}{2\mu} \|\nabla f(x)\|^2.$$

En combinant ces deux dernières inégalités, et en utilisant le fait que $\rho = \alpha/L$ avec $\alpha \in]0, 2[$, on obtient :

$$f(x^+) - \inf f \leq f(x) - \inf f - \frac{1}{2L} \alpha(2-\alpha) \|\nabla f(x)\|^2 \leq (f(x) - \inf f) \left(1 - \frac{\mu}{L} \alpha(2-\alpha)\right).$$

On conclut avec le fait que $\alpha(2-\alpha) \in]0, 1[$. ■

Démonstration du Théorème IV.42 avec le bon θ . La preuve exacte de ce résultat est assez technique, et peut être trouvée dans [17, Theorem 3.3]. ■

A.II.2 Méthode du gradient : cas convexe

Dans le cas fortement convexe on a essentiellement utilisé le fait que l'algorithme est θ -Lipschitzien avec $\theta < 1$. Cela nous permet de prouver facilement par récurrence que les itérés convergent. Lorsque f n'est que convexe, le problème est que l'algorithme devient seulement 1-Lipschitzien :

Lemme A.13 (Non-expansivité de la méthode du gradient). Soit $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$. Soit $\mathbb{A}_\rho : \mathbb{R}^N \rightarrow \mathbb{R}^N$, $x \mapsto x - \rho \nabla f(x)$. Alors, pour tout $\rho \in [0, 2/L]$, \mathbb{A}_ρ est 1-Lipschitzienne :

$$(\forall x, y \in \mathbb{R}^N) \quad \|\mathbb{A}_\rho x - \mathbb{A}_\rho y\| \leq \|x - y\|.$$

Démonstration. Si on regarde la preuve du Théorème IV.37 (dans le Chapitre IV ou dans la Section A.II.1), on voit qu'elle marche encore si $\mu = 0$ (ce qui est notre cas ici) et si $\rho \in [0, 2/L]$. On en déduit donc que pour tout $\rho \in [0, 2/L]$, \mathbb{A}_ρ est θ -Lipschitzienne, avec

$$\theta = \max\{|1 - \rho 0|; |1 - \rho L|\} = \max\{1; |1 - \rho L|\} = 1. \quad \blacksquare$$

Donc tout ce que l'on peut dire c'est que

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\|. \quad (\text{A.6})$$

Or le fait que cette suite soit décroissante ne veut pas dire qu'elle tend vers 0. Il va donc falloir obtenir des inégalités plus précises pour améliorer (A.6).

A.II.2.i) Convergence des valeurs en $O(1/k)$: pas court $\rho \leq 1/L$

Lemme A.14 (Variations de la distance aux solutions). Soient $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ et $x^* \in \operatorname{argmin} f$. Soient $x \in \mathbb{R}^N$ quelconque et $x_+ := x - \rho \nabla f(x)$, avec $\rho L \in]0, 2[$. Alors :

$$\|x_+ - x^*\|^2 - \|x - x^*\|^2 \leq (L\rho - 1) \|x_+ - x\|^2 - 2\rho(f(x_+) - \inf f).$$

Démonstration. On calcule :

$$\begin{aligned}
 & \frac{1}{2\rho} \|x_+ - x^*\|^2 - \frac{1}{2\rho} \|x - x^*\|^2 \\
 = & -\frac{1}{2\rho} \|x - x_+\|^2 - \frac{1}{\rho} \langle x - x_+, x_+ - x^* \rangle \quad \text{en développant les carrés} \\
 = & -\frac{1}{2\rho} \|x - x_+\|^2 - \langle \nabla f(x), x_+ - x^* \rangle \quad \text{d'après la définition de } x_+ \\
 = & -\frac{1}{2\rho} \|x - x_+\|^2 + \langle \nabla f(x), x^* - x \rangle - \langle \nabla f(x), x_+ - x \rangle \quad \text{en faisant } \pm x. \tag{A.7}
 \end{aligned}$$

D'une part, on sait via la convexité de f et l'inégalité des hyperplans (Proposition III.13.ii)) que

$$\langle \nabla f(x), x^* - x \rangle \leq f(x^*) - f(x)$$

D'autre part on sait via la Lipschitzianité de ∇f et (IV.2) que

$$-\langle \nabla f(x), x_+ - x \rangle \leq \frac{L}{2} \|x_+ - x\|^2 + f(x) - f(x_+).$$

En combinant tout cela on en déduit que

$$\frac{1}{2\rho} \|x_+ - x^*\|^2 - \frac{1}{2\rho} \|x - x^*\|^2 \leq f(x^*) - f(x_+) + \left(\frac{L}{2} - \frac{1}{2\rho} \right) \|x_+ - x\|^2.$$

■

Démonstration du Théorème IV.44 pour un pas court. On suppose ici que $\rho L \in]0, 1]$. L'idée de la preuve va être de montrer qu'une certaine « énergie » décroît au cours des itérations. On connaît déjà deux quantités qui décroissent : $f(x_k) - \inf f$ (cf. Proposition IV.34), ainsi que $\|x_k - x^*\|^2$ (cf. Lemme A.14). Dans cette preuve on va considérer une certaine combinaison de ces deux quantités :

$$E_k := k(f(x_k) - \inf f) + c\|x_k - x^*\|^2, \quad \text{avec } c = \frac{1}{2\rho}. \tag{A.8}$$

Pour montrer que l'énergie E_k décroît, nous allons montrer que sa variation est négative :

$$\begin{aligned}
 & E_{k+1} - E_k \\
 = & (k+1)(f(x_{k+1}) - \inf f) - k(f(x_k) - \inf f) + c\|x_{k+1} - x^*\|^2 - c\|x_k - x^*\|^2 \\
 = & f(x_{k+1}) - \inf f + k(f(x_{k+1}) - f(x_k)) + c(\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2) \\
 \leq & f(x_{k+1}) - \inf f + c(\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2),
 \end{aligned} \tag{A.9}$$

où dans la dernière inégalité on utilise le fait que $f(x_{k+1}) - f(x_k) \leq 0$ (cf. Proposition IV.34). Avec le résultat du Lemme A.14 on obtient

$$E_{k+1} - E_k \leq f(x_{k+1}) - \inf f + c(L\rho - 1)\|x_{k+1} - x_k\|^2 - 2\rho c(f(x_{k+1}) - \inf f).$$

Puisque $\rho L \leq 1$ et $2\rho c = 1$, on conclut que E_k est bien décroissante.

Cela nous permet alors d'écrire que

$$k(f(x_k) - \inf f) \leq E_k \leq E_0 = c\|x_0 - x^*\|^2.$$

En divisant cette inégalité par k , on obtient bien que

$$(\forall k \geq 1) \quad f(x_k) - \inf f \leq \frac{\|x_0 - x^*\|^2}{2\rho k}.$$



A.II.2.ii) Convergence des valeurs en $O(1/k)$: Pas long $\rho \geq 1/L$

Lemme A.15 (Égalité du parallélogramme généralisée). Soient $x, y \in \mathbb{R}^N$ et $\alpha \in \mathbb{R}$. Alors

$$\|(1-\alpha)x + \alpha y\|^2 = (1-\alpha)\|x\|^2 + \alpha\|y\|^2 - \alpha(1-\alpha)\|x-y\|^2.$$

Démonstration. On développe les carrés pour écrire :

$$\|(1-\alpha)x + \alpha y\|^2 = (1-\alpha)^2\|x\|^2 + \alpha^2\|y\|^2 + 2\alpha(1-\alpha)\langle x, y \rangle.$$

Ensuite on utilise le fait que

$$2\langle x, y \rangle = \|x\|^2 + \|y\|^2 - \|x-y\|^2$$

ainsi que le fait que $\alpha^2 + \alpha(1-\alpha) = \alpha$ et $(1-\alpha)^2 + (1-\alpha)\alpha = (1-\alpha)$ pour conclure. ■

Le résultat suivant montre que la méthode du gradient est un peu mieux que 1-Lipschitzienne. C'est un résultat analogue au Lemme V.57 pour la projection.

Lemme A.16 (Non-expansivité de la méthode du gradient : avancé). Soient $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$ et $x^* \in \operatorname{argmin} f$. Soit $\mathbb{A}_\rho : x \mapsto x - \rho \nabla f(x)$, avec $\rho L \in]0, 2]$. Alors :

$$(\forall x, y \in \mathbb{R}^N) \quad \|\mathbb{A}_\rho x - \mathbb{A}_\rho y\|^2 \leq \|x - y\|^2 - \gamma \|(I - \mathbb{A}_\rho)x - (I - \mathbb{A}_\rho)y\|^2, \quad (\text{A.10})$$

avec $\gamma = \frac{2-\rho L}{\rho L}$.

Démonstration. On a vu dans le Lemme A.13 que \mathbb{A}_ρ est 1-Lipschitzienne. On va par la suite utiliser une décomposition astucieuse de \mathbb{A}_ρ (qui se vérifie immédiatement à la main) :

$$\mathbb{A}_\rho = (1 - \alpha)I + \alpha T, \quad \text{où} \quad \alpha = \frac{\rho L}{2} \in]0, 1[\text{ et } T = \mathbb{A}_{2/L}.$$

L'idée est de voir que $\mathbb{A}_{2/L}$ n'est que 1-Lipschitzienne, mais l'identité I est très « gentille », donc si \mathbb{A}_ρ est une combinaison convexe de $\mathbb{A}_{2/L}$ et I , alors \mathbb{A}_ρ devrait être un peu mieux que 1-Lipschitzienne. Cela nous permet d'écrire :

$$\begin{aligned} & \|\mathbb{A}_\rho x - \mathbb{A}_\rho y\|^2 \\ = & \|(1 - \alpha)(x - y) + \alpha(Tx - Ty)\|^2 \quad \text{car } \mathbb{A}_\rho = (1 - \alpha)I + \alpha T \\ = & (1 - \alpha)\|x - y\|^2 + \alpha\|Tx - Ty\|^2 - \alpha(1 - \alpha)\|(I - T)x - (I - T)y\|^2. \quad (\text{Lemme A.15}) \end{aligned}$$

D'une part, on sait que $T = \mathbb{A}_{2/L}$ est 1-Lipschitzienne (Lemme A.13), donc on a

$$(1 - \alpha)\|x - y\|^2 + \alpha\|Tx - Ty\|^2 \leq (1 - \alpha)\|x - y\|^2 + \alpha\|x - y\|^2 = \|x - y\|^2.$$

D'autre part, nous avons par définition de T que $I - T = \frac{1}{\alpha}(I - \mathbb{A}_\rho)$, donc

$$\alpha(1 - \alpha)\|(I - T)x - (I - T)y\|^2 = \frac{(1 - \alpha)}{\alpha}\|(I - \mathbb{A}_\rho)x - (I - \mathbb{A}_\rho)y\|^2.$$

En combinant toutes ces inégalités on conclut que

$$\|\mathbb{A}_\rho x - \mathbb{A}_\rho y\|^2 \leq \|x - y\|^2 - \frac{1 - \alpha}{\alpha}\|(I - \mathbb{A}_\rho)x - (I - \mathbb{A}_\rho)y\|^2.$$

où $\frac{1 - \alpha}{\alpha} = \frac{2 - \rho L}{\rho L}$. ■

Démonstration du Théorème IV.44 pour un pas long. On suppose ici que $\rho L \in [1, 2[$. Ici nous allons considérer la même énergie qu'en (A.8), mais avec une constante différente :

$$E_k := k(f(x_k) - \inf f) + c\|x_k - x^*\|^2, \quad \text{avec} \quad c = \frac{1}{2\rho} \left(1 + \frac{\rho L - 1}{\gamma}\right) > 0, \quad (\text{A.11})$$

où $\gamma > 0$ est la constante apparaissant dans le Lemme A.16. Pour montrer que l'énergie E_k décroît, on commence comme pour le pas court et on obtient la même chose que (A.9) :

$$E_{k+1} - E_k \leq f(x_{k+1}) - \inf f + c \left(\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2 \right). \quad (\text{A.12})$$

Le Lemme A.14 nous dit que

$$\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2 \leq (\rho L - 1)\|x_{k+1} - x_k\|^2 - 2\rho(f(x_{k+1}) - \inf f). \quad (\text{A.13})$$

On a aussi le Lemme A.16, que l'on peut utiliser avec $x = x_k$ et $y = x^*$, en exploitant le fait que $\mathbb{A}_\rho x^* = x^* - \rho \nabla f(x^*) = x^*$, ce qui nous donne :

$$\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2 \leq -\gamma \|x_{k+1} - x_k\|^2. \quad (\text{A.14})$$

Posons $\sigma := \frac{\gamma}{\gamma + \rho L - 1}$. C'est un simple exercice que de vérifier que, puisque $\rho L \in [1, 2[$, alors $\gamma > 0$ et $\sigma \in]0, 1]$. On va donc multiplier (A.13) par σ , et (A.14) par $(1 - \sigma)$, pour obtenir

$$\begin{aligned} & \|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2 \\ & \leq -\sigma 2\rho(f(x_{k+1}) - \inf f) + \sigma(\rho L - 1)\|x_{k+1} - x_k\|^2 - (1 - \sigma)\gamma \|x_{k+1} - x_k\|^2 \\ & = -\sigma 2\rho(f(x_{k+1}) - \inf f) + (\sigma(\rho L - 1) - (1 - \sigma)\gamma) \|x_{k+1} - x_k\|^2. \end{aligned} \quad (\text{A.15})$$

On peut calculer que

$$\sigma(\rho L - 1) - (1 - \sigma)\gamma = \frac{\gamma}{\gamma + \rho L - 1}(\rho L - 1) - \frac{\rho L - 1}{\gamma + \rho L - 1}\gamma = 0,$$

d'où

$$c \left(\|x_{k+1} - x^*\|^2 - \|x_k - x^*\|^2 \right) \leq -c\sigma 2\rho(f(x_{k+1}) - \inf f).$$

Or $c\sigma 2\rho = 1$, donc si on combine cette inégalité avec (A.12), on en déduit que E_k est décroissante. On peut alors conclure, comme pour le pas court :

$$(\forall k \geq 1) \quad f(x_k) - \inf f \leq \frac{c\|x_0 - x^*\|^2}{k},$$

où ici $c = \frac{1}{2\rho} \left(\frac{1+(\rho L-1)^2}{2-\rho L} \right)$. ■

A.II.2.iii) Convergence des itérés

Lemme A.17 (d'Opial). Soit $C \subset \mathbb{R}^N$ convexe fermé non vide. Soit $(x_k)_{k \in \mathbb{N}} \subset \mathbb{R}^N$ une suite telle que :

- 1) pour tout $x^* \in C$, la suite $\|x_k - x^*\|$ converge,
- 2) toute valeur d'adhérence de $(x_k)_{k \in \mathbb{N}}$ appartient à C .

Alors x_k converge vers un $x^* \in C$.

Démonstration. (Voir [16, Lemma 5.2]) D'après i), on sait que la suite x_k est bornée. Donc il existe une sous-suite convergente $x_{n_k} \rightarrow x_\infty$, avec $x_\infty \in C$ d'après ii). Puisque $x_\infty \in C$ on peut utiliser i) pour dire que toute la suite $\|x_n - x_\infty\|^2$ tend vers une limite, notons-là ℓ . Si c'est vrai pour toute la suite, ça l'est aussi pour notre sous-suite : $\|x_{n_k} - x_\infty\|^2 \rightarrow \ell$. Or on sait que $\|x_{n_k} - x_\infty\|^2 \rightarrow 0$; donc $\ell = 0$. D'où $\|x_n - x_\infty\|^2 \rightarrow 0$, et donc x_n converge vers un élément de C . ■

Démonstration du Théorème IV.44 : convergence de x_k . Notons que l'hypothèse du Théorème IV.44 nous dit que $\operatorname{argmin} f$ est non vide. Par ailleurs, puisque f est convexe continue, alors on sait que $\operatorname{argmin} f$ est un ensemble convexe fermé. On va donc pouvoir appliquer le Lemme d'Opial A.17 à notre suite, avec $C = \operatorname{argmin} f$. Pour conclure il nous faut vérifier ses deux hypothèses.

Premièrement, soit $x^* \in \operatorname{argmin} f$, et montrons que la suite $\|x_k - x^*\|$ converge. Avec par exemple (A.14) on voit que cette suite est décroissante, donc elle converge bien. Deuxièmement, supposons qu'il existe une sous-suite x_{k_n} qui converge vers un vecteur x^* . Alors on peut utiliser le fait qu'on a déjà prouvé que $f(x_k)$ converge vers $\inf f$. En particulier, la sous-suite $f(x_{k_n})$ converge aussi vers $\inf f$. Or f est continue, donc $f(x_{k_n})$ converge vers $f(x^*)$. On a donc montré que $f(x^*) = \inf f$, ce qui veut bien dire que $x^* \in \operatorname{argmin} f$. ■

A.II.3 Méthode du gradient projeté : cas convexe

L'opérateur de la méthode du gradient projeté vaut $\mathbb{A} = \operatorname{proj}_C \circ \mathbb{A}_\rho$, où proj_C est l'opérateur de projection sur C , et \mathbb{A}_ρ est l'opérateur correspondant à la méthode du gradient, que l'on a bien étudié dans la section précédente. Sans forte convexité, on ne peut pas espérer beaucoup plus que la 1-Lipschitzianité de \mathbb{A} :

Lemme A.18 (Non-expansivité de la méthode du gradient projeté). Soit $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$. Soit $\mathbb{A} : \mathbb{R}^N \rightarrow \mathbb{R}^N$, $x \mapsto \operatorname{proj}_C(x - \rho \nabla f(x))$. Alors, pour tout $\rho \in [0, 2/L]$, \mathbb{A} est 1-Lipschitzienne :

$$(\forall x, y \in \mathbb{R}^N) \quad \|\mathbb{A}x - \mathbb{A}y\| \leq \|x - y\|.$$

Démonstration. Il suffit d'utiliser le fait que $\mathbb{A} = \operatorname{proj}_C \circ \mathbb{A}_\rho$, où proj_C et \mathbb{A}_ρ sont 1-Lipschitziennes (Théorème V.58 et Lemme A.13). ■

A.II.3.i) Convergence des valeurs en $O(1/k)$: pas court $\rho \leq 1/L$

Pour un pas court on peut obtenir la même estimation que pour la méthode du gradient :

Lemme A.19 (Gradient projeté : Variations de la distance aux solutions). Soient $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$, $C \subset \mathbb{R}^N$ convexe fermé non vide, et $x^* \in \operatorname{argmin}_C f$. Soient $x \in \mathbb{R}^N$ quelconque et $x_+ := \operatorname{proj}_C(x - \rho \nabla f(x))$, avec $\rho L \in]0, 2]$. Alors :

$$\|x_+ - x^*\|^2 - \|x - x^*\|^2 \leq (L\rho - 1)\|x_+ - x\|^2 - 2\rho(f(x_+) - \inf f). \quad (\text{A.16})$$

Démonstration. On va reprendre les mêmes arguments que pour la preuve du Lemme A.14. On développe les carrés pour obtenir

$$\frac{1}{2\rho}\|x_+ - x^*\|^2 - \frac{1}{2\rho}\|x - x^*\|^2 = -\frac{1}{2\rho}\|x - x_+\|^2 - \frac{1}{\rho}\langle x - x_+, x_+ - x^* \rangle.$$

Maintenant il nous faut exprimer comment x et x_+ sont reliés. Pour cela on revient à la définition $x_+ := \text{proj}_C(x - \rho \nabla f(x))$, et on applique la caractérisation de la projection par les angles (Proposition V.53) pour écrire

$$\begin{aligned} & \langle x^* - x_+, (x - \rho \nabla f(x)) - x_+ \rangle \leq 0 \\ \Leftrightarrow & \langle x^* - x_+, x - x_+ \rangle \leq \rho \langle x^* - x_+, \nabla f(x) \rangle \\ \Leftrightarrow & -\frac{1}{\rho} \langle x - x_+, x_+ - x^* \rangle \leq -\langle x_+ - x^*, \nabla f(x) \rangle. \end{aligned}$$

On a donc obtenu la même inégalité qu'en (A.7). On peut donc continuer de la même façon que dans la preuve du Lemme A.14, et conclure. ■

Démonstration du Théorème IV.44 pour un pas court. La preuve est exactement la même que pour la méthode du gradient avec pas court (voir la Section A.II.2). La seule différence est qu'il faudra utiliser les variations du Lemme A.19, et le fait que $f(x_k)$ est décroissante (Proposition V.65). ■

A.II.3.ii) Convergence des valeurs en $O(1/k)$: pas long $\rho \geq 1/L$

Lemme A.20 (Non-expansivité de la méthode du gradient projeté : avancé). Soient $f \in \Gamma_0(\mathbb{R}^N) \cap C_L^{1,1}(\mathbb{R}^N)$, $C \subset \mathbb{R}^N$ convexe fermé non vide, et $x^* \in \text{argmin}_C f$. Soit $\mathbb{A} : x \mapsto \text{proj}_C(x - \rho \nabla f(x))$, avec $\rho L \in]0, 2]$. Alors, avec $\gamma = \frac{2-\rho L}{2}$:

$$(\forall x, y \in \mathbb{R}^N) \quad \|\mathbb{A}x - \mathbb{A}y\|^2 \leq \|x - y\|^2 - \gamma \|(I - \mathbb{A})x - (I - \mathbb{A})y\|^2. \quad (\text{A.17})$$

Démonstration. (Adapté de [2, Proposition 4.44]) On écrit $\mathbb{A} = \text{proj}_C \circ \mathbb{A}_\rho$ où $\mathbb{A}_\rho x = x - \rho \nabla f(x)$. On commence par utiliser successivement les résultats du Lemme V.57 pour proj_C , et du Lemme A.16 pour \mathbb{A}_ρ , pour écrire

$$\begin{aligned} & \|\mathbb{A}x - \mathbb{A}y\|^2 \\ = & \|\text{proj}_C \circ \mathbb{A}_\rho x - \text{proj}_C \circ \mathbb{A}_\rho y\|^2 \\ \leq & \|\mathbb{A}_\rho x - \mathbb{A}_\rho y\|^2 - \|(I - \text{proj}_C) \circ \mathbb{A}_\rho x - (I - \text{proj}_C) \circ \mathbb{A}_\rho y\|^2 \\ \leq & \|x - y\|^2 - \frac{2 - \rho L}{\rho L} \|(I - \mathbb{A}_\rho)x - (I - \mathbb{A}_\rho)y\|^2 - \|(I - \text{proj}_C) \circ \mathbb{A}_\rho x - (I - \text{proj}_C) \circ \mathbb{A}_\rho y\|^2 \end{aligned}$$

Il nous reste à étudier le terme négatif du membre de droite. Pour simplifier les notations, on pose $u = (I - \mathbb{A}_\rho)x - (I - \mathbb{A}_\rho)y$ et $v = (I - \text{proj}_C) \circ \mathbb{A}_\rho x - (I - \text{proj}_C) \circ \mathbb{A}_\rho y$, de telle manière que le terme qui nous intéresse est :

$$\beta \|u\|^2 + \|v\|^2, \text{ où } \beta = \frac{2 - \rho L}{\rho L} > 0.$$

On va normaliser cette quantité en la divisant par $1 + \beta$, afin d'avoir une combinaison convexe, qui nous autorisera à utiliser l'égalité du parallélogramme :

$$\begin{aligned} \frac{\beta}{1+\beta}\|u\|^2 + \frac{1}{1+\beta}\|v\|^2 &= (1-t)\|u\|^2 + t\|v\|^2 \text{ avec } t = \frac{1}{1+\beta} \in [0,1] \\ &= \|(1-t)u + tv\|^2 + t(1-t)\|u - v\|^2 \text{ (Lemme A.15)} \\ &\geq t(1-t)\|u - v\|^2 \\ &= t(1-t)\|(I - \mathbb{A})x - (I - \mathbb{A})y\|^2, \end{aligned}$$

où la dernière égalité vient directement de la définition de u et v , et de la simplification de termes dans le calcul de $u - v$. On a donc prouvé que

$$\|\mathbb{A}x - \mathbb{A}y\|^2 \leq \|x - y\|^2 - (1 + \beta)t(1 - t)\|(I - \mathbb{A})x - (I - \mathbb{A})y\|^2,$$

et on conclut en calculant $(1 + \beta)t(1 - t) = (1 - t) = \frac{2 - \rho L}{2}$. ■

Démonstration du Théorème V.67 pour un pas long. On suppose ici que $\rho L \in [1, 2[$. La preuve est exactement la même que pour la méthode du gradient avec pas long (voir la Section A.II.2). La première différence est qu'on utilisera les Lemmes A.20 et A.19 au lieu des A.16 et A.14. En particulier la valeur de γ va changer, ce qui ne change rien à la preuve, mis à part la valeur de la constante c , qui vaut ici $\frac{L}{2(2 - \rho L)}$. La deuxième différence est que pour une solution $x^* \in \operatorname{argmin}_C f$, on a besoin du fait que $\mathbb{A}x^* = x^*$. Ceci a déjà été vérifié dans la Proposition V.64. ■

A.II.4 Méthode du gradient optimal

Comme pour l'algorithme du gradient à pas fixe (Théorème IV.42, prouvé dans la Section A.II.1) on va ici se contenter de prouver le résultat avec un θ sous-optimal.

Démonstration du Théorème IV.57 avec un θ quelconque. (Voir [14, Eq. (8.47), p.238]) Ici on note ρ_k le pas optimal calculé à l'itération k . D'après le Lemme de Descente (IV.2), on sait que pour tout $\rho > 0$ on a

$$f(x_k - \rho \nabla f(x_k)) \leq f(x_k) + \langle \nabla f(x_k), -\rho \nabla f(x_k) \rangle + \frac{L}{2} \|\rho \nabla f(x_k)\|^2.$$

Si on minimise le terme de gauche par rapport à ρ , on obtient par définition $f(x_{k+1})$. D'un autre côté si on minimise le terme de droite par rapport à ρ , on voit que c'est un polynôme du second degré en ρ . Il est alors facile de voir que le ρ optimal pour le membre de droite est $\rho = \frac{1}{L}$, ce qui nous donne

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L} \|\nabla f(x_k)\|^2.$$

D'autre part on sait d'après la Proposition A.4.i) que

$$f(x_k) - \inf f \leq \frac{1}{2\mu} \|\nabla f(x_k)\|^2.$$

Si on combine ces deux inégalités, on obtient que

$$f(x_{k+1}) - \inf f \leq f(x_k) - \inf f - \frac{1}{2L} \|\nabla f(x_k)\|^2 \leq (1 - \frac{\mu}{L})(f(x_k) - \inf f).$$

D'où le résultat avec $\theta = 1 - \frac{\mu}{L}$. ■

Démonstration du Théorème IV.57 avec le bon θ . Voir [6, Theorem 1.2]. ■

Bibliographie

- [1] G. ALLAIRE, *Analyse Numérique et Optimisation : Une Introduction à La Modélisation Mathématique et à La Simulation Numérique*, Editions Ecole Polytechnique, 2005.
- [2] H. H. BAUSCHKE AND P. L. COMBETTES, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer, 2nd edition ed., 2017.
- [3] V. BECK, J. MALICK, AND G. PEYRÉ, *Objectif Agrégation*, H&K, 2004.
- [4] D. P. BERTSEKAS, *Nonlinear Programming*, Athena Scientific, 1995.
- [5] P. G. CIARLET, *Introduction à l'analyse numérique matricielle et à l'optimisation - 5ème édition*, Dunod, Paris, 2007.
- [6] E. DE KLERK, F. GLINEUR, AND A. B. TAYLOR, *On the worst-case complexity of the gradient method with exact line search for smooth strongly convex functions*, Optimization Letters, 11 (2017), pp. 1185–1199.
- [7] J.-B. HIRIART-URRUTY, *Optimisation et analyse convexe : Exercices et problèmes corrigés, avec rappels de cours*, EDP Sciences, Ulis, France, 2009.
- [8] J.-B. HIRIART-URRUTY AND C. LEMARECHAL, *Convex Analysis and Minimization Algorithms I : Part 1 : Fundamentals*, Springer Science & Business Media, 1996.
- [9] W. KARUSH, *Minima of functions of several variables with inequalities as side constraints*, M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago, (1939).
- [10] T. H. KJELDSEN, *A contextualized historical analysis of the Kuhn–Tucker theorem in nonlinear programming : The impact of World War II*, Historia mathematica, 27 (2000), pp. 331–361.
- [11] N. KOLKIN, J. SALAVON, AND G. SHAKHNAROVICH, *Style transfer by relaxed optimal transport and self-similarity*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10051–10060.
- [12] H. W. KUHN AND A. W. TUCKER, *Nonlinear Programming*, in Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, The Regents of the University of California, 1951.

- [13] J.-L. LAGRANGE, *Manière plus simple et plus générale de faire usage de la formule de l'équilibre donnée dans la section deuxième*, in Mécanique Analytique, vol. 1, 1788, pp. 77–112.
- [14] D. G. LUENBERGER AND Y. YE, *Linear and Nonlinear Programming*, vol. 2, Springer, 1984.
- [15] Y. NESTEROV, *Introductory Lectures on Convex Optimization*, vol. 87, Springer Science & Business Media, 2004.
- [16] J. PEYPOUQUET, *Convex Optimization in Normed Spaces*, SpringerBriefs in Optimization, Springer International Publishing, Cham, 2015.
- [17] A. B. TAYLOR, J. M. HENDRICKX, AND F. GLINEUR, *Exact Worst-Case Convergence Rates of the Proximal Gradient Method for Composite Convex Minimization*, Journal of Optimization Theory and Applications, 178 (2018), pp. 455–476.