OPTIMISATION DÉTERMINISTE
ING2 MATHS-INFO CYTECH

LOUIS GARRIGUE

Soit un ensemble $K \subset \mathbb{R}^d$ une application $f : K \to \mathbb{R}$ bornée inférieurement, c'est-à-dire qu'il existe $c \in \mathbb{R}$ tel que $f(x) \geqslant c$ pour tout $x \in K$. Optimiser f, c'est trouver le ou les minimums globaux de f sur k, c'est-à-dire trouver les solutions du problème

$$\inf_{x \in K} f(x),$$

s'il en existe. Ce problème très général a un ensemble indénombrable d'applications dans tous les domaines de l'ingénierie et des sciences, en intelligence artificielle, en statistiques, en finance, en physique, etc.

Nous ne nous concentrerons pas seulement sur des problèmes abstraits et théoriques, mais nous donnerons les concepts suffisants pour le travail d'un ingénieur, et nous présenterons les méthodes permettant de résoudre très concrètement ces problèmes. Les TPs associés auront de nombreuses applications numériques. Un bon compris entre théorie et application sera suivi pour une formation ingénieur.

Pour réviser et aller plus loin dans le cours, nous conseillons la lecture du livre

"Introduction à l'Optimisation", Jean-Christophe Culioli,
éditions Ellipses [1], Chapitres 1, 2 et 3

Contents

Date: November 11, 2025.

1. Rappels sur la différentiation

Nous aurons besoin de manipuler les dérivées. En effet, par exemple si $f(x) = 3x^2 - x + 1$ et si on veut la minimiser sur $\mathbb{R}$, il faut d'abord trouver les points $x \in \mathbb{R}$ tels que $f'(x) = 0$. Sur $\mathbb{R}^d$, la différentiation est plus délicate.

Généralement, on nomme "application" un élément qui à chaque élément d'un espace de départ, associe un élément d'un espace d'arrivée. Quand l'espace d'arrivée est $\mathbb{R}$, on parle aussi de "fonction".

Dans toute cette section, $\Omega \subset \mathbb{R}^d$ sera ouvert de $\mathbb{R}^d$.

1.1. **Dérivations d'ordre 1.**

1.1.1. *Gradient.* Rappelons que le gradient d'une application $f : \Omega \to \mathbb{R}$, $\mathcal{C}^1$ est l'application notée $\nabla f : \Omega \to \mathbb{R}^d$ donnée par

$$\forall x = (x_1, \dots, x_d) \in \Omega, \qquad (\nabla f)(x) := \begin{pmatrix} (\partial_{x_1} f)(x) \\ \vdots \\ (\partial_{x_d} f)(x) \end{pmatrix}. \tag{1}$$

Par exemple si $d = 2$ et $f(x) = x_1^2 + 3x_2$, alors $(\nabla f)(x) = \begin{pmatrix} 2x_1 \\ 3 \end{pmatrix}$.

Nous noterons aussi $\nabla f(x) := (\nabla f)(x)$ pour alléger le texte.

1.1.2. *Différentielle.* Prenons une fonction $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}^n$, donc plus générale que pour la définition du gradient. On rappelle la définition de la différentielle de f, notée $\mathrm{d}f : \Omega \to \mathcal{L}(\mathbb{R}^d, \mathbb{R}^n)$, où $\mathcal{L}(\mathbb{R}^d, \mathbb{R}^n)$ est l'ensemble des applications linéaires allant de $\mathbb{R}^d$ and $\mathbb{R}^n$.

Definition 1.1 (Différentielle en x)**.** *Prenons $x \in \Omega$. Comme Ω est ouvert, il existe $r > 0$ tel que $B(x, r) \subset \Omega$. Supposons qu'il existe un application linéaire $L : \mathbb{R}^d \to \mathbb{R}^n$ et $c > 0$ et $s \in [0, r]$ tels que pour tout $y \in \mathbb{R}^d$ tel que $\|y\| \leqslant s$,*

$$\|f(x + y) - (f(x) + Ly)\| \leqslant c \|y\|^2 .$$

Alors f est différentiable en x et on note $\mathrm{d}_x f = L$.

On dit que f est différentiable sur Ω si elle l'est en tout point de Ω.

La différentielle généralise les développements limités d'ordre 1, c'est-à-dire qu'on peut écrire que pour tout $x \in \Omega, y \in \mathbb{R}^d$ tel que $x + y \in \Omega$,

$$f(x + y) = f(x) + (\mathrm{d}_x f)\, y + O(\|y\|^2), \tag{2}$$

quand $\|y\| \to 0$. Pour un $x \in \Omega$ donné, $\mathrm{d}_x f$ est une application linéaire allant de $\mathbb{R}^d$ dans $\mathbb{R}^n$. Si f et g sont deux applications composables et $\mathcal{C}^1$, $g \circ f$ est aussi $\mathcal{C}^1$ et la règle de la chaîne donne

$$\mathrm{d}_x(f \circ g)y = \left(\mathrm{d}_{g(x)} f\right) \circ (\mathrm{d}_x g)\, y. \tag{3}$$

Exercice 1.2. *L'application exponentielle* $\exp : \mathcal{M}_d(\mathbb{R}) \to \mathcal{M}_d(\mathbb{R})$ *est définie par* $\exp(A) := \sum_{k=0}^{+\infty} \frac{A^k}{k!}$. *Calculer sa différentielle en 0.*

Solution. On prend $A = 0$. On a

$$\exp(A + H) = \exp H = 1 + H + \sum_{k=2}^{+\infty} \frac{H^k}{k!} = \exp A + H + \sum_{k=2}^{+\infty} \frac{H^k}{k!}.$$

On conjecture que $(\mathrm{d}_A \exp) H = H$. On pose $LH := H$, qui est une application linéaire. On a

$$\begin{aligned} \|\exp(A + H) - ((\exp A) + LH)\| = \left\| \sum_{k=2}^{+\infty} \frac{H^k}{k!} \right\| &\leqslant \sum_{k=2}^{+\infty} \frac{\|H\|^k}{k!} = \|H\|^2 \sum_{k=0}^{+\infty} \frac{\|H\|^k}{(k+2)!} \\ &\leqslant \|H\|^2 \sum_{k=0}^{+\infty} \frac{\|H\|^k}{k!} = \|H\|^2 e^{\|H\|}. \end{aligned}$$

On peut donc écrire

$$\exp(A+H) = \exp(A) + LH + O_{\|H\|\to 0}\left(\|H\|^2\right),$$

donc exp est différentiable en 0 et $(\mathrm{d}_0 \exp)H = LH = H$, i.e. $\mathrm{d}_0 \exp = \mathbb{1}_{\mathcal{M}_d(\mathbb{R})}$. □

1.1.3. *Dérivée.* Prenons un intervalle I. Pour les applications $\xi : I \to \mathbb{R}^d$, c'est-à-dire quand l'espace de départ est plongé dans $\mathbb{R}$, on définit la dérivée de ξ par

$$\xi'(t) := \frac{(\mathrm{d}_t \xi)\, s}{s} \in \mathbb{R}^d$$

pour tout $t \in I$ et tout $s \in \mathbb{R}$. Si $\xi : I \to \mathbb{R}$, la dérivée coïncide avec le gradient.

1.1.4. *Autre définition du gradient.* Reprenons une fonction $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}$. Sur tout point $x \in \Omega$, le gradient $\nabla f(x)$ peut être en fait défini à partir de la différentielle par dualité (via le théorème de Riesz disant que les formes linéaires on un vecteur représentant) via

$$\forall y \in \mathbb{R}^d, \qquad (\mathrm{d}_x f)\, y = \langle \nabla f(x), y \rangle, \tag{4}$$

et on peut montrer qu'il est bien égal à celui donné en (1). La différentielle ne dépend pas du produit scalaire alors que le gradient si.

Lorsque ces quelques règles sont bien maitrisées, on ne se trompe plus dans les calculs de dérivation de fonctions définies sur $\mathbb{R}^d$.

1.1.5. *Dérivée directionnelle.* La dérivée directionnelle généralise en dimension d la dérivée à gauche et à droite de la dimension 1. Dans cette définition, K n'est pas forcément un ensemble ouvert, c'est seulement un sous-ensemble de $\mathbb{R}^d$.

Definition 1.3 (Dérivée directionnelle)**.** *Prenons $K \subset \mathbb{R}^d$, $x \in K$ et $y \in \mathbb{R}^d$. Si la limite*

$$\left(\delta_x^+ f\right)(y) := \lim_{t \to 0^+} \frac{f(x+ty) - f(x)}{t}$$

existe, alors on dit que c'est la dérivée de f en x dans la direction y. On parle aussi de dérivée de Dini.

Par exemple pour $f : [0,1] \to \mathbb{R}$ définie par $f(x) = x$, f n'est pas dérivable en 0 mais a une dérivée de Dini dans la direction 1 (c'est-à-dire à droite) en 0, $\left(\delta_0^+ f\right)(1) = 1$. Un des intérêts de la dérivée directionnelle est qu'elle peut être définie sur le bord ∂K. Si f est différentiable en x, alors pour tout $y \in \mathbb{R}^d$, f a une dérivée de Dini en x dans la direction y et

$$\left(\delta_x^+ f\right)(y) = \langle \nabla f(x), y \rangle. \tag{5}$$

Mais si f a une dérivée en x dans toute direction $y \in \mathbb{R}^d$, f n'est pas forcément différentiable, un contre-exemple est $x \mapsto |x|$ en 0, qui a une dérivée à gauche et à droite mais n'est pas dérivable.

1.2. **Dérivation d'ordre 2.** Soit $f : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ deux fois différentiable. On note $\mathcal{S}_d(\mathbb{R})$ l'ensemble des matrices carrées réelles symétriques de taille $d \times d$. On appelle Hessienne de f l'application $\nabla^{\otimes 2} f : \Omega \to \mathcal{S}_d(\mathbb{R})$, où

$$\left(\nabla^{\otimes 2} f\right)(x) := \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\right)_{1 \leqslant i,j \leqslant d}.$$

Lemma 1.4. *Pour tout $x \in \Omega$, on a bien $\nabla^{\otimes 2} f(x) \in \mathcal{S}_d(\mathbb{R})$.*

Proof. Pour tout $i, j \in \{1, \dots, d\}$, $\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial^2 f}{\partial x_j \partial x_i}(x)$, par le théorème de Schwartz. □

Comme $\nabla^{\otimes 2} f(x)$ est symétrique, on peut la diagonaliser. Il existe une matrice orthogonale $P(x) \in O_d(\mathbb{R})$ (c'est-à-dire que $P(x)P(x)^T = 1$) et $D(x) \in \mathcal{M}_d(\mathbb{R})$ diagonale telles que

$$\nabla^{\otimes 2} f(x) = P(x)D(x)P(x)^{-1}. \tag{6}$$

1.3. **Développement de Taylor.** On a, pour $\|y\| \to 0$, le développement limité d'ordre 2 suivant, pour toute direction $y \in \mathbb{R}^d$,

$$\boxed{f(x+y) = f(x) + \langle \nabla f(x), y \rangle + \tfrac{1}{2} \left\langle y, \left(\nabla^{\otimes 2} f\right) y \right\rangle + O_{\|y\| \to 0}\left(\|y\|^3\right).} \tag{7}$$

1.4. **Exemples de calculs.** Donnons quelques exemples simples.

1.4.1. $t \mapsto f(ta + b)$. Supposons $\Omega = \mathbb{R}^d$ et $a, b \in \mathbb{R}^d$, et prenons $f : \mathbb{R} \to \mathbb{R}$. On définit $w : \mathbb{R} \to \mathbb{R}$ par

$$w(t) := f(ta + b)$$

et on veut calculer sa dérivée. On peut poser $m : \mathbb{R} \to \mathbb{R}^d$ définie par $m(t) := ta + b$, on a alors $w = f \circ m$, donc la règle de la chaîne (3) donne, pour tout $t, s \in \mathbb{R}$,

$$(\mathrm{d}_t w)\, s = \left(\mathrm{d}_{m(t)} f\right) \circ (\mathrm{d}_t m)\, s.$$

Or, on peut calculer que $(\mathrm{d}_t m)s = as$ donc

$$\begin{aligned} w'(t) &= \frac{\left(\mathrm{d}_{m(t)} f\right) \circ (\mathrm{d}_t m)\, s}{s} = \frac{\left(\mathrm{d}_{m(t)} f\right) \circ as}{s} = \left(\mathrm{d}_{m(t)} f\right) \circ a = \langle \nabla f(m(t)), a \rangle \\ &= \langle \nabla f(ta + b), a \rangle. \end{aligned}$$

Le calcul est plus simple qu'en passant par (1). De même, on peut calculer

$$w''(t) = \left\langle a, \left(\left(\nabla^{\otimes 2} f\right)(ta + b)\right) a \right\rangle.$$

1.4.2. *Norme.* Calculons la différentielle de l'application $f : \mathbb{R}^d \to \mathbb{R}$ définie par $f(x) := \|x\|^2$.

On a $f(x) = f(x_1, \dots, x_d) = \sum_{k=1}^d x_k^2$ donc

$$\left(\partial_{x_j} f\right)(x) = \sum_{k=1}^d \partial_{x_j} x_k^2 = \sum_{k=1}^d 2\delta_{k=j} x_k = 2x_j$$

et

$$\nabla f(x) = \begin{pmatrix} (\partial_{x_1} f)(x) \\ \vdots \\ (\partial_{x_d} f)(x) \end{pmatrix} = \begin{pmatrix} 2x_1 \\ \vdots \\ 2x_d \end{pmatrix} = 2x.$$

On calcule $\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \sum_{k=1}^d \frac{\partial^2}{\partial x_i \partial x_j} x_k^2 = 2\delta_{i=j}$ donc $\nabla^{\otimes 2} f(x) = 2 \times \mathbb{1} = 2$ où $\mathbb{1}$ est la matrice identité.

1.4.3. *Formes quadratiques.*

Proposition 1.1: Gradient et Hessienne d'une forme quadratique

Soit $b \in \mathbb{R}^d$ et $c \in \mathbb{R}$, $A \in \mathcal{M}_n(\mathbb{R})$ et $f : \mathbb{R}^d \to \mathbb{R}$ telle que $\forall x \in \mathbb{R}^d$,

$$f(x) := \langle Ax, x\rangle + \langle b, x\rangle + c. \tag{8}$$

On a alors

$$\nabla f(x) = \left(A^T + A\right) x + b, \qquad \text{et} \qquad \nabla^{\otimes 2} f(x) = A^T + A.$$

Sous cette forme, f est appelée une forme quadratique.

Proof. On fait un calcul direct et un calcul via la différentielle. On verra que le calcul utilisant la différentielle est beaucoup plus simple et rapide.

- Calcul direct. On appelle $g(x) := \langle Ax, x\rangle$ et $h(x) := \langle b, x\rangle$. Alors $g(x) = \sum_{i=1}^d (Ax)_i x_i = \sum_{1\leqslant i,j\leqslant d} A_{ij} x_j x_i$ donc

$$\begin{aligned}
\left(\partial_{x_k} g\right)(x) &= \sum_{1\leqslant i,j\leqslant d} A_{ij} \partial_{x_k} x_j x_i = \sum_{1\leqslant i,j\leqslant d} A_{ij} \left(2x_k \delta_{k=i=j} + x_i \delta_{k=j, i\neq k} + x_j \delta_{k=i, j\neq k}\right) \\
&= 2A_{kk} x_k + \sum_{\substack{1\leqslant i\leqslant d \\ i\neq k}} A_{ik} x_i + \sum_{\substack{1\leqslant j\leqslant d \\ j\neq k}} A_{kj} x_j = \sum_{i=1}^d A_{ik} x_i + \sum_{j=1}^d A_{kj} x_j \\
&= \left(A^t x\right)_k + (Ax)_k .
\end{aligned}$$

De même, $h(x) = \sum_{i=1}^d b_i x_i$ donc $\partial_{x_k} h(x) = b_k$.

- En utilisant $\mathrm{d}_x f$. On a

$$\begin{aligned}
f(x+y) &= \langle A(x+y), x+y\rangle + \langle b, x+y\rangle + c \\
&= f(x) + \langle Ax, y\rangle + \langle Ay, x\rangle + \langle Ay, y\rangle + \langle b, y\rangle \\
&= f(x) + \left\langle (A + A^T)x + b, y\right\rangle + O(\|y\|^2).
\end{aligned}$$

Donc pour que la définition (2) soit respectée, on déduit la différentielle $(\mathrm{d}_x f)y = \left\langle (A + A^T)x + b, y\right\rangle$. Or, le gradient est défini via (4) donc $\nabla f(x) = (A + A^T)x + b$.

- Hessienne. On calcule

$$\frac{\partial^2 g}{\partial x_i \partial x_j}(x) = \sum_{1\leqslant k,p\leqslant d} A_{kp} \frac{\partial^2}{\partial x_i \partial x_j} x_k x_p = \sum_{\substack{1\leqslant k,p\leqslant d \\ \text{ou} \\ k=j, p=i}} A_{kp} \delta_{k=i, p=j} = A_{ij} + A_{ji} = (A + A^T)_{ij}.$$

donc $\nabla^{\otimes 2} f(x) = A + A^T$. □

1.4.4. *Forme quadratique particulière.* Si $f(x_1, x_2) := x_1^2 + \lambda x_2^2$ où $\alpha \in \mathbb{R}$, alors $\nabla f(x) = 2\begin{pmatrix} x_1 \\ \lambda x_2 \end{pmatrix}$ and $\nabla^{\otimes 2} f(x) = 2\begin{pmatrix} 1 & 0 \\ 0 & \lambda \end{pmatrix}$.

1.5. **Rappels sur les matrices.**

Definition 1.5 (Matrices symétriques)**.** *Prenons $M \in \mathcal{M}_d(\mathbb{R})$.*

- *Si $M^T = M$, on dit que M est symétrique et on note $M \in \mathcal{S}_d(\mathbb{R})$*
- *Si $M \in \mathcal{S}_d(\mathbb{R})$ et si toutes les valeurs propres de M sont positives, on dit que M est positive et on note $M \in \mathcal{S}_d^+(\mathbb{R})$ ou $M \geqslant 0$, c'est équivalent à ce que $\forall y \in \mathbb{R}^d, \langle y, My\rangle \geqslant 0$.*
- *Si $M \in \mathcal{S}_d(\mathbb{R})$ et si toutes les valeurs propres de M sont strictement positives, on dit que M est définie positive et on note $M \in \mathcal{S}_d^{++}(\mathbb{R})$ ou $M > 0$, c'est équivalent à ce que $\forall y \in \mathbb{R}^d, \langle y, My\rangle > 0$.*

Pour M symétrique, on a $M = P^T D P$ où $D = \operatorname{diag}\left(\lambda_1, \dots, \lambda_d\right)$ et P est orthogonale, donc pour tout $y \in \mathbb{R}^d$,

$$\langle y, My\rangle = \left\langle y, P^T DPy\right\rangle = \langle Py, DPy\rangle = \sum_{j=1}^{d} \lambda_j \left((Py)_j\right)^2. \tag{9}$$

Preuve dans la définition 1.5 . Montrons que $M \in \mathcal{S}_d^+(\mathbb{R})$ est équivalent à $\forall y \in \mathbb{R}^d, \langle y, My\rangle \geqslant 0$.

Supposons que toutes les valeurs propres $\lambda_1, \dots, \lambda_d$ sont positives. Soit $y \in \mathbb{R}^d$, alors on voit par (9) que $\langle y, My\rangle \geqslant 0$.

Supposons que pour tout $y \in \mathbb{R}^d$, $\langle y, My\rangle \geqslant 0$. Soit $e_1, \dots, e_d \in \mathbb{R}^d$ la base canonique de $\mathbb{R}^d$. Soit $j \in \{1, \dots, d\}$, alors on voit par (9) que $\langle P^{-1}e_j, MP^{-1}e_j\rangle = \lambda_j$ et donc $\lambda_j \geqslant 0$. □

Partie algorithmique

2. Algorithme de descente générique

Numériquement, minimiser f sur K consiste souvent, par un algorithme, à démarrer d'un élément $x_0 \in K$ et à générer une suite $(x_n)_{n\in\mathbb{N}} \in K^{\mathbb{N}}$ telle que

$$f(x_{n+1}) \leqslant f(x_n).$$

On espère que cette suite converge vers un minimum global.

2.1. **Descente.** L'intérêt principal de la définition de la dérivée de Dini 1.3 est que

- Si $\left(\delta_x^+ f\right)(y) > 0$ alors f est croissante dans la direction y,
- Si $\left(\delta_x^+ f\right)(y) < 0$ alors f est décroissante dans la direction y.

Une notion encore plus faible est la direction de descente.

Definition 2.1 (Direction de descente)**.** *Soient $f : \mathbb{R}^d \to \mathbb{R}$ et $x \in \mathbb{R}^d$. Le vecteur $y \in \mathbb{R}^d$ est une direction de descente pour f à partir du point x si $t \mapsto f(x + ty)$ est décroissante en $t = 0$, c'est-à-dire s'il existe $\eta > 0$ tel que $\forall t \in]0, \eta]$, $f(x+ty) < f(x)$.*

Si f est dérivable en $x \in K$ dans la direction $y \in \mathbb{R}^d$ et que $(\delta_x^+ f)(y) < 0$, alors y est une direction de descente. Mais y peut être une direction de descente sans que la dérivée de f en x dans la direction y soit définie. Penser par exemple à une fonction discontinue, $f : \mathbb{R} \to \mathbb{R}$, $f(t) := 0$ si $t \neq 0$ et $f(0) := 1$, alors f n'est pas dérivable en 0 à droite ou à gauche mais -1 et 1 sont des directions de descente au point $t = 0$.

La dérivée directionnelle permet de savoir si une direction est de descente.

Proposition 2.2. *Soient $f : K \to \mathbb{R}$ dérivable en $x \in K$ dans la direction $y \in \mathbb{R}^d$. Si $(\delta_x^+ f)(y) < 0$ alors y est une direction de descente. Si y est une direction de descente alors $(\delta_x^+ f)(y) \leqslant 0$.*

Proof. On a le développement

$$f(x + ty) = f(x) + t\left(\delta_x^+ f\right)(y) + O_{t\to 0}\left(t^2\right).$$

Si $(\delta_x^+ f)(y) < 0$, alors $(\delta_x^+ f)(y) \neq 0$ et pour $t > 0$ petit, $f(x + ty) - f(x)$ a le signe de $(\delta_x^+ f)(y)$. Si y est une direction de descente, alors pour $t > 0$ assez petit $f(x + ty) - f(x) \leqslant 0$ et on déduit que $(\delta_x^+ f)(y) \leqslant 0$. □

En particulier, grâce à (5), le gradient permet de savoir si on est sur une direction de descente. Mais il permet encore mieux, il permet de donner la direction de plus forte descente. L'ensemble des directions possibles est donné par $\mathbb{S} := \{y \in \mathbb{R}^d, \|y\| = 1\}$, on ne retient pas la norme car c'est la direction qui importe.

Proposition 2.3 (Direction de plus forte descente)**.** *Soit $f : K \to \mathbb{R}$ une fonction différentiable sur $\mathring{K}$ et $x \in \mathring{K}$. Alors le problème*

$$\min_{\substack{y\in\mathbb{R}^d \\ \|y\|=1}} \langle \nabla f(x), y\rangle \tag{10}$$

est minimisé pour $y_* = -\frac{\nabla f(x)}{\|\nabla f(x)\|}$, qui est la direction de plus forte descente.

Proof. Le problème (10) est posé sur l'ensemble compact $\mathbb{S}$ et la fonction $y \mapsto \langle \nabla f(x), y\rangle$ est continue car linéaire, donc le problème admet un minimiseur global.

Pour tout $y \in \mathbb{S}$, on a

$$\langle \nabla f(x), y_*\rangle = -\|\nabla f(x)\| = -\|\nabla f(x)\|\,\|y\| \underset{\substack{\text{Cauchy-}\\ \text{Schwartz}}}{\leqslant} \langle \nabla f(x), y\rangle.$$

□

2.2. **Algorithme générique.** Afin de générer une suite $(x_n)_{n\in\mathbb{N}}$ telle que $f(x_{n+1}) \leqslant f(x_n)$, on va chercher une direction de descente et itérer. Le principe suivant sera suivi par les algorithmes que nous utiliserons. On considère qu'on dispose de fonctions indépendantes permettant d'évaluer f et ∇f en n'importe quel point $x \in K$.

Le choix du pas peut ne pas nécessiter de faire appel à f, dans ce cas l'algorithme demande seulement des appels à ∇f et la fonction sera de la forme `descent(`∇f`,` x_0`)`.

Algorithm 1 Algorithme générique de descente

procedure DESCENT(f, ∇f, x_0)
  $k \leftarrow 0$
  while test de convergence non satisfait **do**
    Trouver une direction $y_k \in \mathbb{R}^d$ telle que $\langle \nabla f(x_k), y_k \rangle < 0$ (i.e. direction de descente)
    Choisir un pas $s_k > 0$ tel que $f(x_k + s_k y_k) < f(x_k)$
    $x_{k+1} = x_k + s_k y_k$
    $k \leftarrow k + 1$
  end while
  return x
end procedure

2.3. **Test de convergence / arrêt de l'algorithme.** On privilégie les erreurs relatives aux erreurs absolues, voir Appendice A pour la définition de la distance relative.

Soit x_* un point de minimum local de f. Le test d'arrêt idéal serait $x_k = x_*$ mais bien sûr on ne connaît pas x_*. En pratique, on choisit une tolérance $\varepsilon \in \mathbb{R}_+^*$ et on adopte un ou plusieurs critères d'arrêt parmi les suivants

- Optimalité : $\|\nabla f(x_k)\| < \varepsilon$,
- Stagnation de la solution : $\mathbb{D}(x_{k+1}, x_k) < \varepsilon$,
- Stagnation de la valeur : $\mathbb{D}(f(x_{k+1}), f(x_k)) < \varepsilon$,
- Nombre d'itérations k dépassant un seuil fixé.

2.4. **Convergence.**

Definition 2.4 (Algorithme globalement convergent)**.** *Soit un algorithme itératif qui génère une suite $(x_k)_{k\in\mathbb{N}}$ pour résoudre $\min_{x\in K} f(x)$, où f est $\mathcal{C}^1(K)$. L'algorithme est dit globalement convergent si, quel que soit le point initial $x_0 \in K$, $\|\nabla f(x_k)\| \underset{k\to+\infty}{\longrightarrow} 0$.*

Attention, cette notion ne garantit pas que la limite soit un minimum, même local. L'algorithme peut converger vers un point critique qui n'est pas un minimum, un point selle par exemple.

2.5. **Choix du pas.** Considérons que nous avons une méthode pour choisir la direction de descente y_k. Il reste maintenant à définir une stratégie de recherche linéaire pour le calcul du pas s_k. Nous donnons ici deux manières naturelles.

2.5.1. *Pas constant.* Il consiste simplement à prendre $s_k = s$ constant pour tout $k \in \mathbb{N}$. La question est alors comment choisir un pas qui garantisse la convergence de l'algorithme ? Un pas trop petit aidera à la convergence mais il faudra beaucoup d'itérations alors qu'un pas trop grand ne fera pas converger.

2.5.2. *Pas optimal.* On peut choisir un pas qui rende la fonction à minimiser la plus petite possible dans cette direction. Cette méthode est appelée méthode à pas

optimal. Il reste donc, à chaque itération de l'algorithme 1, à résoudre le problème

$$\inf_{t\in\mathbb{R}} f\left(x_k + sy_k\right),$$

et on appelle $s_k = s$ un minimiseur global s'il existe.

Lemma 2.5 (Pas optimal pour les formes quadratiques)**.** *Soit $A \in \mathcal{S}_d^{++}(\mathbb{R})$ et $b \in \mathbb{R}^d$, on considère la forme quadratique $Q(x) := \frac{1}{2}\langle x, Ax\rangle - \langle b, x\rangle$. Imaginons qu'on itère avec une formule $x^{k+1} = x^k + sy^k$, c'est-à-dire dans la direction y^k. Alors le pas optimal s^k est*

$$s^k = \frac{\left\langle b - Ax^k, y^k\right\rangle}{\left\langle y^k, Ay^k\right\rangle}. \tag{11}$$

Proof. On définit

$$g(s) := Q(x^k + sy^k) = Q(x^k) + s\left\langle x^k, Ay^k\right\rangle - s\left\langle b, y^k\right\rangle + \frac{1}{2}s^2\left\langle y^k, Ay^k\right\rangle.$$

On a

$$g'(s) = \left\langle x^k, Ay^k\right\rangle - \left\langle b, y^k\right\rangle + s\left\langle y^k, Ay^k\right\rangle,$$

et $g''(s) = \left\langle y^k, Ay^k\right\rangle > 0$. g est donc fortement convexe et a donc un unique minimiseur global. On le trouve en résolvant $g'(s^k) = 0$, qui donne le résultat. □

La résolution de ce problème de minimisation unidimensionnel peut coûter cher en temps de calcul si on doit le trouver numériquement. Pour cette raison, on peut lui préférer parfois l'algorithme de gradient à pas constant.

3. Vitesse de convergence

Soit $(x_k)_{k\in\mathbb{N}} \subset \mathbb{R}^{\mathbb{N}}$ une suite convergente donnée par un algorithme, on note $x_* := \lim_{k\to+\infty} x_k$.

Dans la plupart des cas rencontrés usuellement, les vitesses de convergence des algorithmes sont telles que

$$\frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|^\gamma} \underset{k\to+\infty}{\longrightarrow} \alpha.$$

pour $\alpha > 0$ et $\gamma \geqslant 1$. Nous ne considérerons pas d'autres cas.

Definition 3.1 (Vitesse de convergence)**.** *Supposons qu'il existe $\alpha > 0$ et $\gamma \geqslant 1$ tels que*

$$\frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|^\gamma} \underset{k\to+\infty}{\longrightarrow} \alpha.$$

- *Si $\gamma = 1$ et $\alpha < 1$, on dit que la convergence est linéaire et α est appelé taux de convergence*
- *Si $\gamma = \alpha = 1$, nous référons à la Section 3.1*
- *Si $\gamma > 1$ on dit que la convergence est superlinéaire d'ordre γ. Si $\gamma = 2$, on dit que la convergence est quadratique.*

Nous ne regarderons $\gamma = 1$ et $\alpha > 1$ car alors $\|x_k - x_*\| \to +\infty$.

3.1. **Cas $\gamma = \alpha = 1$.** Dans le cas où $\gamma = \alpha = 1$, on ne sait même pas *a priori* si x_k converge, par exemple $x_k = k$ et $x_k = 1/k$ sont dans cette situation. Nous pourrons essayer de conjecturer quelque chose de plus précis, comme dans la définition suivante.

Definition 3.2 (Conjecture fréquente lorsque $\gamma = \alpha = 1$)**.** *Si $\gamma = \alpha = 1$, nous aurons souvent*

$$\|x_k - x_*\| \underset{k\to+\infty}{\sim} \frac{C}{k^p}, \tag{12}$$

où $C > 0$ et $p > 0$.

Si on a (12), on voit que

$$\frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} = \left(1 + \frac{1}{k}\right)^{-p} \underset{k\to+\infty}{\longrightarrow} 1,$$

ce qui correspond à la grandeur de la Définition 3.1 avec $\gamma = \alpha = 1$.

3.2. **Remarque sur le cas $\gamma = 1$ et $\alpha < 1$.**

Proposition 3.1: Convergence dans le cas $\gamma = 1$ et $\alpha < 1$

Supposons que

$$\frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|} \underset{k\to+\infty}{\longrightarrow} \alpha.$$

où $\alpha < 1$, alors il existe $\beta \in [\alpha, 1[$ et $C >$ tel que pour k assez grand,

$$\|x_k - x_*\| \leqslant C\beta^k.$$

Si en plus $\alpha > 0$, alors $\|x_k - x_*\| \geqslant C'(\beta')^k$ pour $C' > 0$, $\beta' \in]0, \alpha]$ et k assez grand.

Proof. Posons $S_k := \|x_k - x_*\|$. Dans ce cas, il existe $N \in \mathbb{N}$ tel que pour tout $k \geqslant N$, $\alpha - \varepsilon \leqslant \frac{S_{k+1}}{S_k} \leqslant \alpha + \varepsilon$, où $\alpha + \varepsilon < 1$. Pour tout $k \geqslant N$ on a alors

$$(\alpha - \varepsilon)^{k-N} S_N \leqslant S_k \leqslant (\alpha + \varepsilon)^{k-N} S_N$$

et donc en posant $C_0 := \alpha^{-N} S_N$, $C_- := (1 - \varepsilon/\alpha)^{-N}$ et $C_+ := (1 + \varepsilon/\alpha)^{-N}$, on a

$$C_- \left(1 - \frac{\varepsilon}{\alpha}\right)^k \leqslant \frac{S_k}{C_0\alpha^k} \leqslant C_+ \left(1 + \frac{\varepsilon}{\alpha}\right)^k,$$

On ne peut pas conclure que $\lim_{k\to+\infty} \frac{S_k}{C_0\alpha^k} = 1$ sans ajouter d'autres conditions, mais on peut intuitivement penser que $S_k \simeq C_0\alpha^k$ (pas au sens des équivalents). En fait en définissant $\beta := \alpha + \varepsilon < 1$ et $C := C_0C_+$, on a $S_k \leqslant C\beta^k$. On peut faire pareil pour la borne inférieure. □

Si on a exactement $\frac{\|x_{k+1}-x_*\|}{\|x_k-x_*\|} = \alpha$, alors $S_k = S_0\alpha^k$. Dans tous les cas, on a une convergence exponentielle de $\|x_k - x_*\|$ vers 0, même si on parle de vitesse de convergence linéaire !

3.3. **Remarque sur le cas $\gamma > 1$.** Dans le cas où $\gamma > 1$, en supposant que $\frac{\|x_{k+1}-x_*\|}{\|x_k-x_*\|^\gamma} = \alpha$ exactement, posons $S_k := \|x_k - x_*\|$. Nous avons alors

$$S_k = \alpha S_{k-1}^\gamma = \alpha^{1+\gamma} S_{k-2}^{\gamma^2} = \alpha^{1+\gamma+\gamma^2} S_{k-3}^{\gamma^3} = \cdots = \alpha^{1+\gamma+\gamma^2+\cdots+\gamma^{k-1}} S_0^{(\gamma^k)} = \alpha^{\frac{\gamma^k-1}{\gamma-1}} S_0^{(\gamma^k)}$$
$$= e^{\gamma^k\left((\gamma-1)^{-1}\ln\alpha+\ln S_0\right)+(\gamma-1)^{-1}\ln\alpha}.$$

On voit que $S_k \to 0$ est équivalent à $(\gamma-1)^{-1}\ln\alpha + \ln S_0 < 0$, lui-même équivalent à $S_0 < \alpha^{\frac{1}{1-\gamma}}$. Par exemple si $\alpha = \gamma = 2$, il faut que $S_0 < 1/4$. Ceci montre qu'intuitivement, il faudra partir avec u_0 pas trop loin de la solution u pour que l'algorithme converge, sinon il divergera. Dans le cas de convergence vers 0, c'est une convergence super-exponentielle de $\|x_k - x_*\|$ vers 0.

Proposition 3.2: Convergence dans le cas $\gamma > 1$

Supposons que

$$\frac{\|x_{k+1} - x_*\|}{\|x_k - x_*\|^\gamma} \underset{k\to+\infty}{\longrightarrow} \alpha.$$

où $\gamma > 1$, alors il existe $\beta > 0$ et $C >$ tels que pour k assez grand,

$$\|x_k - x_*\| \leqslant C\beta^{\gamma^k}.$$

Attention ce résultat ne précise pas si $\beta < 1$, qui est nécessaire pour avoir la convergence vers 0.

Proof. En reprenant le calcul précédent, on a, pour N assez grand, $S_{k+1}/S_k^\gamma \leqslant \alpha+\varepsilon$ où $\alpha+\varepsilon < 1$ (on considère que c'est vrai pour $N \geqslant 0$ sans perte de généralité),

$$S_k \leqslant (\alpha+\varepsilon)S_{k-1}^\gamma \leqslant \ldots \leqslant (\alpha+\varepsilon)^{\frac{\gamma^k-1}{\gamma-1}} S_0^{(\gamma^k)} = C(\alpha+\varepsilon)^{a\gamma^k}$$

où $C := (\alpha+\varepsilon)^{(\gamma-1)^{-1}}$ et $a := (\gamma-1)^{-1} + \ln S_0$, il reste à choisir $\beta := (\alpha+\varepsilon)^a$. □

3.4. **Calculer les constantes en pratique.** Pour évaluer le type de vitesse de convergence donné en Définition 3.2, on pourra commencer par calculer numériquement γ puis α. En définissant $S_k := \|x_k - x_*\|$ et $J_k := S_{k+1}/S_k$, on pourra évaluer numériquement les limites

$$\lim_{k\to+\infty} \frac{\ln J_{k+1}}{\ln J_k} = \gamma, \qquad \text{puis} \qquad \lim_{k\to+\infty} \frac{S_{k+1}}{S_k^\gamma} = \alpha.$$

Si $\alpha = 1$, on pourra conjectuer un comportement asymptotique de type Définition 3.2. On pourra alors numériquement calculer

$$p = -\lim_{k\to+\infty} \left(\ln(1+k^{-1})\right)^{-1} \ln J_k = -\lim_{k\to+\infty} k \ln J_k, \tag{13}$$

et enfin

$$C = \lim_{k\to+\infty} \frac{S_k}{k^p}.$$

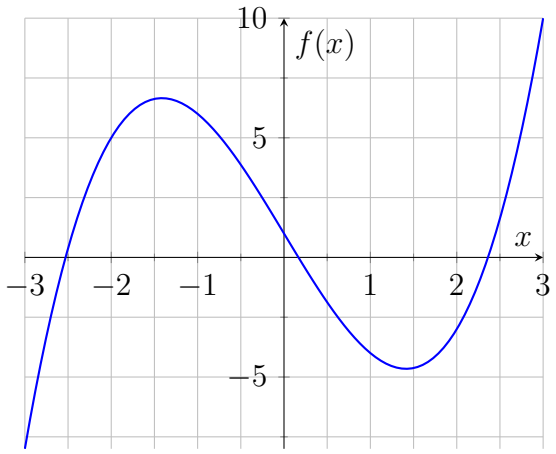


FIGURE 1. Graph de la fonction $f(x) = x^3 - 6x + 1$

3.5. **Exemple de convergence.** Considérons $f(x) = x^3 - 6x + 1$, tracée en Figure 1. L'unique minimum local est en $x_* = \sqrt{2}$. Partant de $x_0 = 2$, on compare plusieurs algorithmes

- $x_{k+1} = x_k - \mu(x_k^2 - 2)$. Pour $0 < \mu < 1/\sqrt{2}$, convergence linéaire avec taux $\alpha = \left|2\mu\sqrt{2} - 1\right|$. Si $\mu = (2\sqrt{2})^{-1}$, la convergence est superlinéaire.
- $x_{k+1} = \frac{1}{2}\left(x_k + \frac{2}{x_k}\right)$. Convergence quadratique : en théorie 4 itérations suffisent pour obtenir 5 chiffres exacts. En pratique, on a 11 chiffres significatifs après 4 itérations.

4. QUELQUES ALGORITHMES DE DESCENTE

Un algorithme de descente est déterminé par les stratégies de choix des directions de descente successives y_k, et des pas s_k dans cette direction. Nous avons déjà vu comment obtenir des pas en Section 2.5.

4.1. **Descente de gradient.** Nous suivons le schéma général de l'algorithme 1. Le choix de la direction de plus forte descente $y_k = -\nabla f(x_k)$ donnée par la Proposition 2.3 définit une famille d'algorithmes appelés algorithmes de descente de gradient. Le pas peut ensuite être choisi comme en Section 2.5.

4.1.1. *Gradient à pas fixe.* L'itération est alors

$$x^{k+1} = x^k - \mu\nabla f(x^k), \quad \mu > 0,$$

où $\mu > 0$ est un paramètre constant qu'on doit régler manuellement. Si μ est choisi trop grand, l'algorithme diverge et si μ est trop petit, la convergence est très lente. Lorsqu'on ajoute d'autres hypothèses, on peut prouver la convergence linéaire.

Théorème 4.1: Convergence linéaire du gradient à pas fixe

Supposons f différentiable, a-fortement convexe, et $x \mapsto \nabla f(x)$ est L-Lipschiztienne. Alors pour tout $\mu \in]0, \frac{2}{L}[$ l'algorithme de gradient à pas μ fixe converge vers l'unique minimum, la vitesse de convergence étant linéaire. Le plus petit rapport de convergence est $\sqrt{1 - \frac{a^2}{L^2}}$ et il est atteint pour $\mu = \frac{a}{L^2}$.

Proof. Notons $e^k := x^k - x_*$ où x_* est l'unique minimiseur de f. Alors $e^{k+1} = e^k - \mu \nabla f(x^k)$ et

$$\begin{aligned}\left\|e^{k+1}\right\|^2 &= \left\|e^k\right\|^2 - 2\mu \left\langle \nabla f(x^k), e^k \right\rangle + \mu^2 \left\|\nabla f(x^k)\right\|^2 \\ &\underset{\nabla f(x_*)=0}{=} \left\|e^k\right\|^2 - 2\mu \left\langle \nabla f(x^k) - \nabla f(x^*), e^k \right\rangle + \mu^2 \left\|\nabla f(x^k) - \nabla f(x^*)\right\|^2 \\ &\leqslant \left\|e^k\right\|^2 (1 - 2\mu a + \mu^2 L^2).\end{aligned}$$

On définit $P(\mu) := \mu^2 L^2 - 2\mu a + 1$ pour tout $\mu \geqslant 0$. On a que le minimum de P est atteint pour $\mu = \frac{a}{L^2}$, et $P\left(\frac{a}{L^2}\right) = 1 - \frac{a^2}{L^2}$. Il faut avoir $L > a$ pour que ce minimum soit positif. Dans ce cas $P(\mu) \geqslant 0$ pour tout $\mu > 0$, et on a, pour tout $\mu > 0$,

$$\frac{\left\|e^{k+1}\right\|}{\left\|e^k\right\|} \leqslant \sqrt{P(\mu)},$$

La méthode converge tant que $\sqrt{P(\mu)} < 1$, c'est-à-dire tant que $\mu < \frac{2a}{L^2}$. La vitesse de convergence est alors linéaire de rapport $\sqrt{P(\mu)}$, qui est minimisé pour $\mu = \frac{a}{L^2}$. □

Le gradient à pas fixe est simple mais ses performances sont assez mauvaises.

4.1.2. *Gradient à pas optimal.*

Definition 4.1 (Conditionnement d'une matrice)**.** *Soit $A \in \mathcal{S}_d^{++}(\mathbb{R})$, on note $0 < \lambda_1 \leqslant \cdots \leqslant \lambda_d$ les valeurs propres de A en ordre croissant. Le conditionnement de A relativement à la norme euclidienne est*

$$\kappa := \frac{\lambda_d}{\lambda_1}.$$

On voit que $\kappa \geqslant 1$.

On définit la forme bilinéaire $\langle x, y \rangle_A := \langle x, Ay \rangle$, qui est bien un produit scalaire quand $A \in \mathcal{S}^{++}(\mathbb{R})$.

L'algorithme du gradient à pas optimal permet de résoudre l'équation $Ax = b$.

Lemma 4.2 (Convergence du gradient à pas optimal pour les formes quadratiques)**.** *Soit $A \in \mathcal{S}_d^{++}(\mathbb{R})$ et $b \in \mathbb{R}^d$. Considérons la forme quadratique $Q(x) := \frac{1}{2} \langle x, Ax \rangle - \langle b, x \rangle$. La méthode du gradient à pas optimal converge vers l'optimum unique $x^* = A^{-1}b$ à vitesse linéaire*

$$\left\|x^{k+1} - x_*\right\|_A \leqslant \frac{\kappa - 1}{\kappa + 1} \left\|x^k - x_*\right\|_A.$$

Le rapport de convergence est $\alpha = \frac{\kappa-1}{\kappa+1} \in]0,1[$. Pour que ce rapport soit le plus petit possible, il faut que κ soit le plus proche de 1 possible, c'est-à-dire qu'il faut que toutes les valeurs propres de A soient à peu près égales. Si c'est le cas, on dit alors que la matrice A est bien conditionnée, c'est la situation voulue puisque la convergence est alors plus rapide.

Proof. Pour simplifier la suite, nous allons noter $G^k := \nabla f(x^k)$. Nous réutilisons le Lemme 2.5, dans lequel la direction est $y^k = G^k$ car nous faisons une descente de gradient. On a $b - Ax^k = G^k$ donc le pas optimal est

$$\alpha^k = \frac{\left\|G^k\right\|^2}{\left\|G^k\right\|_A^2}. \tag{14}$$

Pour tout k,

$$e^{k+1} = e^k - \alpha^k G^k.$$

Notons que A est symétrique définie positive. Elle admet donc une racine carrée B, c'est-à-dire telle $A = B^T B$ (par exemple, la factorisation de Cholesky donne $A = CC^T$; choisir $B = C^T$). Donc $Be^{k+1} = Be^k - \alpha^k BG^k$. On prend le carré, ce qui donne

$$\begin{aligned}
\left\|e^{k+1}\right\|_A^2 = \left\|Be^{k+1}\right\|^2 &= \left\|Be^k\right\|^2 + (\alpha^k)^2 \left\|BG^k\right\|^2 - 2\alpha^k \left\langle Be^k, BG^k\right\rangle \\
&\underset{Ae^k=G^k}{=} \left\|e^k\right\|_A^2 - 2\alpha^k \left\|G^k\right\|^2 + (\alpha^k)^2 \left\|G^k\right\|_A^2 \\
&\underset{(14)}{=} \left\|e^k\right\|_A^2 - 2\frac{\left\|G^k\right\|^2}{\left\|G^k\right\|_A^2} \left\|G^k\right\|^2 + \left(\frac{\left\|G^k\right\|^2}{\left\|G^k\right\|_A^2}\right)^2 \left\|G^k\right\|_A^2 = \left\|e^k\right\|_A^2 - \frac{\left\|G^k\right\|^4}{\left\|G^k\right\|_A^2} \\
&\underset{\left\|e^k\right\|_A = \left\|G^k\right\|_{A^{-1}}}{=} \left\|e^k\right\|_A^2 \left(1 - \frac{\left\|G^k\right\|^4}{\left\|G^k\right\|_A^2 \left\|G^k\right\|_{A^{-1}}^2}\right).
\end{aligned}$$

Nous aurons besoin du lemme suivant.

Lemma 4.3 (Inégalité de Kantorovitch)**.** *Soit $A \in \mathcal{S}^{++}(\mathbb{R})$, de valeurs propres λ_i (positives) ordonnées dans le sens suivant $0 < \lambda_1 \leqslant \cdots \leqslant \lambda_d$. Alors pour tout $x \in \mathbb{R}^d$,*

$$\frac{4\lambda_1\lambda_d}{(\lambda_1+\lambda_d)^2} \leqslant \frac{\|x\|^4}{\|x\|_A^2 \|x\|_{A^{-1}}^2}. \tag{15}$$

Nous donnons une preuve de ce lemme en Appendice B.1. On applique l'inégalité de Kantorovich (15),

$$\|e^{k+1}\|_A^2 \leqslant \|e^k\|_A^2 \left(1 - \frac{4\lambda_1\lambda_n}{(\lambda_1+\lambda_n)^2}\right) = \|e^k\|_A^2 \left(\frac{\lambda_n-\lambda_1}{\lambda_n+\lambda_1}\right)^2 = \|e^k\|_A^2 \left(\frac{\kappa-1}{\kappa+1}\right)^2.$$

□

On constate donc, d'après ce théorème, que lorsque la matrice A est mal conditionnée (i.e. quand κ est grand), l'algorithme de gradient à pas optimal converge très lentement.

Voici la version plus générale.

Theorem 4.4 (Convergence du gradient à pas optimal)**.** *Soit f une fonction $\mathcal{C}^1$ et a-fortement convexe. On considère l'algorithme de gradient à pas optimal. Alors*

- *pour tout $k \in \mathbb{N}$, $\langle \nabla f(x_k), \nabla f(x_{k+1}) \rangle = 0$,*
- *$x_k \to x_*$ quand $k \to +\infty$.*

Quand on fait une descente de gradient à pas optimal, on peut constater graphiquement $\nabla f(x_k) \perp \nabla f(x_{k+1})$. Ceci crée des zigzags qui ne produisent pas une trajectoire directe et efficace.

Proof. Comme f est a-fortement convexe, elle admet un unique minimiseur x_*. On définit

$$\varphi(s) := f(x_k - s\nabla f(x_k)), \qquad \varphi'(s) = -\langle \nabla f(x_k), \nabla f(x_k - s\nabla f(x_k)) \rangle .$$

L'algorithme de descente de gradient à pas optimal fait que le pas optimal minimise φ, le minimiseur étant s_k, qui vérifie $\varphi'(s_k) = 0$, ce qui se réécrit $\langle \nabla f(x_k), \nabla f(x_{k+1}) \rangle = 0$. Ceci implique que $\nabla f(x^{k+1}) \perp x^{k+1} - x^k$. Comme f est fortement convexe de rapport a, on en déduit que

$$f(x^k) \geqslant f(x^{k+1}) + \left\langle \nabla f(x^k), x^k - x^{k+1} \right\rangle + \frac{a}{2} \left\| x^{k+1} - x^k \right\|^2 = f(x^{k+1}) + \frac{a}{2} \left\| x^{k+1} - x^k \right\|^2 .$$

La suite $(f(x^k))_{k\in\mathbb{N}}$ est donc décroissante, d'autre part elle est minorée par $f(x^*)$, elle converge donc. On en déduit que $f(x^k) - f(x^{k+1}) \to 0$, soit encore $\|x^{k+1} - x^k\| \to 0$. Puisque $(f(x^k))_{k\in\mathbb{N}}$ est bornée et que f est coercive, alors $(x^k)_{k\in\mathbb{N}}$ est bornée, $x^k \in B_R(0)$ pour un certain $R > 0$. Comme $x \mapsto \nabla f(x)$ est continue, elle est uniformément lipschitzienne sur $B_R(0)$, et on déduit que $\nabla f(x^k) - \nabla f(x^{k+1}) \to 0$. Or

$$\|\nabla f(x^k)\|^2 = \left\langle \nabla f(x^k), \nabla f(x^k) - \nabla f(x^{k+1}) \right\rangle ,$$

ce qui implique que $\|\nabla f(x^k)\| \to 0$. Enfin, grâce à la forte convexité,

$$\frac{a}{2}\|x^k - x_*\|^2 \leqslant \left\langle \nabla f(x^k) - \nabla f(x_*), x^k - x_* \right\rangle \leqslant \|\nabla f(x^k) - \nabla f(x^*)\| \|x^k - x_*\|.$$

On utilise que $\nabla f(x^*) = 0$, alors

$$\|x^k - x_*\| \leqslant \frac{2}{a}\|\nabla f(x^k)\|.$$

□

4.2. **Méthode de Newton locale.** Pour construire les méthodes de gradient, nous avons remplacé f par son approximation linéaire au voisinage de l'itéré courant. Or, elles ne sont pas très performantes, en partie parce qu'elles ne tiennent pas compte de la courbure (ou de la Hessienne) qui est une information de second ordre. Nous prenons cette information en compte ici.

Supposons que f est $\mathcal{C}^2$ et définissons la forme quadratique correspondant à l'ordre de 2 de f au voisinage de l'itéré courant x_k

$$q(y) := f(x_k) + \langle \nabla f(x_k), y - x_k \rangle + \tfrac{1}{2} \left\langle y - x_k, \nabla^{\otimes 2} f(x_k)(y - x_k) \right\rangle .$$

On a

$$f(y) = q(y) + O_{\|y - x_k\| \to 0} \left(\|y - x_k\|^3 \right) .$$

Si $\nabla^{\otimes 2} f(x_k)$ est définie positive, alors q est strictement convexe et coercive, et elle admet un unique minimiseur global x_*. Celui-ci vérifie l'équation d'Euler $\nabla q(x_*) = 0$, équivalente à

$$\nabla f(x_k) + \nabla^{\otimes 2} f(x_k)(x_* - x_k) = 0,$$

c'est-à-dire $x_* = x_k - \nabla^{\otimes 2} f(x_k)^{-1} \nabla f(x_k)$. On choisit alors le point x_{k+1} comme si on avait exactement $f = q$, c'est-à-dire $x_{k+1} = x_* = x_k - \nabla^{\otimes 2} f(x_k)^{-1} \nabla f(x_k)$. On reconnaît ici l'algorithme de Newton de recherche de zéro de la fonction ∇f. La méthode ne doit jamais être appliquée en utilisant une inversion de la Hessienne (qui peut être de très grande taille et mal conditionnée) mais plutôt en utilisant

$$\boxed{x_{k+1} = x_k + y_k} \tag{16}$$

où y_k est l'unique solution du système linéaire

$$\boxed{\nabla^{\otimes 2} f(x_k) y_k = -\nabla f(x_k)} \tag{17}$$

y_k est appelée direction de Newton. Le critère d'arrêt sera $\|\nabla f(x_k)\| < \varepsilon$.

Cette méthode est bien définie si à chaque itération, la matrice hessienne $\nabla^{\otimes 2} f(x_k)$ est définie positive. Ceci est vrai en particulier au voisinage de la solution x_* cherchée si on suppose que $\nabla^{\otimes 2} f(x_*)$ est définie positive, par continuité de $\nabla^{\otimes 2} f$.

La méthode de Newton est un algorithme de descente à pas fixe égal à 1 et de direction $y_k = -\nabla^{\otimes 2} f(x_k)^{-1} \nabla f(x_k)$. Si la fonctionnelle f est quadratique, strictement convexe, alors l'algorithme converge en une seule itération.

4.2.1. *Convergence.* L'algorithme hérite des propriétés de l'algorithme de Newton.

Théorème 4.2: Convergence quadratique la méthode de Newton

Soit f de classe $\mathcal{C}^3$ et x_* un minimiseur local de f. On suppose que $\nabla^{\otimes 2} f(x_*)$ est définie positive. Il existe un voisinage $\mathcal{V}$ de x_* tel que si $x_0 \in \mathcal{V}$, alors la suite des itérés $(x_k)_{k\in\mathbb{N}}$ générés à partir de x_0 par la méthode de Newton locale, converge vers x_*. De plus, la convergence est au moins quadratique.

Proof. On définit l'erreur $e_k = x_k - x_*$. Le développement de Taylor de ∇f autour de x_* donne

$$\nabla f(x_k) = \nabla f(x_*) + \nabla^{\otimes 2} f(x_*)(x_k - x_*) + O\left(\|e_k\|^2\right) = \nabla^{\otimes 2} f(x_*) e_k + O\left(\|e_k\|^2\right).$$

L'itération est $x_{k+1} = x_k - \left(\nabla^{\otimes 2} f(x_k)\right)^{-1} \nabla f(x_k)$, et pour x_k proche de x_*, on a $\nabla^{\otimes 2} f(x_k) = \nabla^{\otimes 2} f(x_*) + O\left(\|e_k\|\right)$, et comme $\nabla^{\otimes 2} f(x_*)$ est inversible et $f \in \mathcal{C}^3$, $\nabla^{\otimes 2} f(x)$ est aussi inversible pour tout x dans un voisinage de x_*, et $\left(\nabla^{\otimes 2} f(x_k)\right)^{-1} = \left(\nabla^{\otimes 2} f(x_*)\right)^{-1} + O\left(\|e_k\|\right)$. On estime

$$\begin{aligned} e_{k+1} &= e_k - \left(\nabla^{\otimes 2} f(x_k)\right)^{-1} \nabla f(x_k) \\ &= e_k - \left(\left(\nabla^{\otimes 2} f(x_*)\right)^{-1} + O\left(\|e_k\|\right)\right)\left(\nabla^{\otimes 2} f(x_*) e_k + O\left(\|e_k\|^2\right)\right) \\ &= O\left(\|e_k\|^2\right). \end{aligned}$$

On a donc que $\frac{\|e_{k+1}\|}{\|e_k\|^2}$ est borné, donc la convergence est au moins quadratique. □

La méthode peut diverger si le point initial n'est pas suffisamment proche d'un point de minimum local, et elle n'est pas définie si les matrices $\nabla^{\otimes 2} f(x_k)$ ne sont pas définies positives. Utilisée dans le cadre de l'optimisation, la méthode de Newton locale présente un autre inconvénient : la solution identifiée à la fin de l'algorithme n'est pas forcément un point de minimum local, mais uniquement un point critique de f.

4.3. **Gradient conjugué.** L'algorithme du gradient conjugué permet de résoudre des systèmes linéaires

$$Ax = b$$

où A est strictement positive de grande taille. Il peut servir à

- résoudre des systèmes linéaires, utiles dans (17) par exemple,
- être un modèle de base pour la résolution itérée de problèmes d'optimisation non linéaire comme on le verra avec l'algorithme 3.

4.3.1. *Résoudre le système linéaire.* Résoudre le système linéaire $Ax = b$ avec $A \in \mathcal{M}_d(\mathbb{R})$ symétrique définie positive, $b \in \mathbb{R}^d$, où l'inconnue est $x \in \mathbb{R}^d$, est équivalent à minimiser la forme quadratique

$$Q(x) = \tfrac{1}{2} \langle x, Ax \rangle - \langle b, x \rangle .$$

Ces deux problèmes ont effectivement une même solution unique. En tout point x qui n'est pas cette solution, on note l'erreur

$$r(x) := Ax - b = \nabla Q(x),$$

appelé résidu, qui est non nul.

Definition 4.5 (Directions conjuguées)**.** *Un ensemble $\{p^0, p^1, ..., p^{\ell-1}\}$ de vecteurs de $\mathbb{R}^d$ est dit conjugué par rapport à A si $\forall i, j \in \{0, \dots, \ell - 1\}$, $i \neq j$,*

$$\left\langle p^i, Ap^j \right\rangle = 0. \tag{18}$$

Lemma 4.6. *Soit A définie positive. Une famille conjuguée par rapport à A est libre.*

Proof. Supposons qu'il existe $(\alpha_j)_{0 \leqslant j \leqslant \ell-1} \in \mathbb{R}^\ell \backslash \{0\}$ tel que $\sum_{j=0}^{\ell-1} \alpha_j p^j = 0$. Alors en appliquant A à gauche et en prenant le produit scalaire avec p^i, on obtient

$$0 = \sum_{j=0}^{\ell-1} \left\langle p^i, Ap^j \right\rangle = \left\langle p^i, Ap^i \right\rangle ,$$

ce qui contredit le fait que A est définie positive. □

Un exemple d'une telle famille est donné par une famille de vecteurs propres orthogonaux de A.

4.3.2. *Minimisation suivant les directions conjuguées.* Dans l'algorithme de gradient conjugué en question,

$$x^{k+1} = x^k + s^k p^k \tag{19}$$

les directions successives pour l'optimisation sont selon les directions conjuguées p^k. On cherche le pas optimal pour ces directions. Il est donné par le Lemme 2.5,

$$s^k = -\frac{\langle r^k, p^k \rangle}{\langle p^k, Ap^k \rangle}, \tag{20}$$

où $r^k = r(x^k)$.

Theorem 4.7. *Soit $x^0 \in \mathbb{R}^d$, et on considère la séquence générée par (19), où s^k est donné par (20) et où $\{p^0, p^1, ..., p^{d-1}\}$ est une famille de d directions conjuguée par rapport à A, avec $A \in \mathcal{M}_d(\mathbb{R})$ symétrique définie positive. Alors*

- *la solution x_* de $Ax = b$ est atteinte en au plus d étapes.*
- $\langle r^k, p^i \rangle = 0 \quad \forall k \in \{0, ..., d-1\}, \forall i \in \{0, ..., k-1\}$
- *x^k est le minimiseur global de Q sur l'espace affine*

$$\mathcal{A} := \{x^0 + \textstyle\sum_{j=0}^{k-1} \mu^j p^j \mid (\mu^0, ..., \mu^{k-1}) \in \mathbb{R}^k\}.$$

Proof.

• Les directions $p^0, p^1, ..., p^{d-1}$ sont linéairement indépendantes donc elles engendrent $\mathbb{R}^d$. Il existe donc $\sigma^0, ..., \sigma^{d-1} \in \mathbb{R}$ tels que

$$x_* - x^0 = \sigma^0 p^0 + ... + \sigma^{d-1} p^{d-1}. \tag{21}$$

En prenant le produit scalaire avec Ap^k, pour tout $k = 0, ..., d-1$, on a

$$\sigma^k = \frac{\langle Ap^k, x_* - x^0 \rangle}{\langle p^k, Ap^k \rangle}.$$

Nous allons maintenant montrer que $\sigma^k = s^k$. Après k étapes on a

$$x^k = x^0 + s^0 p^0 + ... + s^{k-1} p^{k-1} \tag{22}$$

qui est tel que $\langle p^k, A(x^k - x^0) \rangle = 0$. Ainsi,

$$\begin{aligned} \langle p^k, A(x_* - x^0) \rangle &= \langle p^k, A(x_* - x^0 - x^k + x^0) \rangle = \langle p^k, A(x_* - x^k) \rangle = \langle p^k, b - Ax^k \rangle \\ &= -\langle p^k, r^k \rangle \end{aligned}$$

d'où $\sigma^k = s^k$. En réutilisant (21) et (22), on obtient $x^d = x_*$. On a en fait trouvé la décomposition de x_* dans la base $p^0, p^1, ..., p^{d-1}$ en minimisant successivement suivant les directions conjuguées de la fonction quadratique Q.

• Prouvons la seconde partie du théorème par récurrence. Le premier point x^1 calculé par la séquence $k = 1$ est obtenu en calculant

$$s^0 = \underset{0 \leqslant s \leqslant d-1}{\operatorname{argmin}} Q(x^0 + sp^0)$$

On a donc $\langle r^1, p^0 \rangle = \langle \nabla Q(x^1), p^0 \rangle = 0$.

Soit $k \in \{2, \ldots, d-1\}$ tel que $\langle r^{k-1}, p^i \rangle = 0, \forall i = 0, ..., k-2$. Calculons maintenant $\langle r^k, p^i \rangle$ pour tout $i = 0, ..., k-1$.

$$r^k = Ax^k - b = A(x^{k-1} + s^{k-1} p^{k-1}) - b = r^{k-1} + s^{k-1} Ap^{k-1} \tag{23}$$

d'où

$$\left\langle r^k, p^{k-1}\right\rangle = \left\langle r^{k-1}, p^{k-1}\right\rangle + s^{k-1}\left\langle r^{k-1}, Ap^{k-1}\right\rangle \underset{(20)}{=} 0.$$

D'autre part, pour tout $i \in \{0, ..., k-2\}$,

$$\left\langle r^k, p^i\right\rangle = \left\langle r^{k-1}, p^i\right\rangle + s^{k-1}\left\langle p^{k-1}, Ap^i\right\rangle = 0,$$

où nous avons utilisé l'hypothèse de récurrence et (18).

• $\mathcal{A}$ est un ensemble convexe et A est strictement convexe donc il existe un unique minimiseur global, qui est aussi un minimiseur local, on applique en fait le théorème 6.5. Pour le trouver, il suffit d'expliciter le gradient de l'application $\mathbb{R}^k \ni s \mapsto Q(x^0 + s^0p^0 + ... + s^{k-1}p^{k-1})$. Sa i-ème coordonnée vaut $\left\langle \nabla Q(x^k), p^i\right\rangle = \left\langle r^k, p^i\right\rangle = 0$ quel que soit $i = 0, ...k-1$, donc x^k vérifie l'inéquation d'Euler et est le minimiseur cherché. □

4.3.3. *Calcul des directions conjuguées.* Il reste à choisir la famille des p^k. On peut construire une base de vecteurs propres orthogonaux car A est symétrique. Ces vecteurs sont par construction conjugués par rapport à A et constituent un choix possible de directions conjuguées. Mais le calcul de ces vecteurs est en général très coûteux en temps de calculs. On peut calculer les directions conjuguées au fur et à mesure qu'on en a besoin par $p^0 := -r^0$ et la récurrence

$$p^k = -r^k + \beta^k p^{k-1} \tag{24}$$

qui s'interprète comme l'opposé du gradient au nouveau point altéré par la direction précédente. Pour que p^k et p^{k-1} soient des directions conjuguées par rapport à A, i.e. pour que $\left\langle p^k, Ap^{k-1}\right\rangle = 0$, on choisit

$$\beta^k = \frac{\left\langle r^k, Ap^{k-1}\right\rangle}{\left\langle r^{k-1}, Ap^{k-1}\right\rangle} \tag{25}$$

La direction p^k est également conjuguée par rapport à A avec les directions p^i pour $i = 0, .., k-2$ comme on va le montrer.

Proposition 4.8. *Avec les notations précédentes, si $x^k \neq x_*$ alors les propriétés suivantes sont vérifiées*

- $\left\langle r^k, r^i\right\rangle = 0 \quad \forall i \in \{0, ..., k-1\}$
- $\mathrm{Vect}(r^0, r^1, ..., r^k) = \mathrm{Vect}(p^0, p^1, ..., p^k) = \mathrm{Vect}(r^0, Ar^0, ..., A^k r^0)$
- $\left\langle p^k, Ap^i\right\rangle = 0 \quad \forall i \in \{0, ..., k-1\}$

Tout espace de la forme $\mathrm{Vect}(u, Au, \ldots, A^k u)$ est appelé espace de Krylov.

Proof.

• Grâce à (24), on a $\mathrm{Vect}(r^0, \ldots, r^i) \subset \mathrm{Vect}(p^0, \ldots, p^i)$ pour tout $i \in \{0, \ldots, k\}$. On peut ensuite montrer par récurrence que

$$\forall i \in \{0, \ldots, d-1\}, \qquad \mathrm{Vect}(r^0, \ldots, r^i) = \mathrm{Vect}(p^0, \ldots, p^i).$$

Or, on sait que $r^k \perp p^j$ pour tout $j \in \{0, \ldots, k-1\}$ donc $r^k \perp \mathrm{Vect}(p^0, \ldots, p^{k-1})$ et on en déduit la première partie.

• La première égalité a déjà été démontrée. On rappelle (23) étant

$$r^k = r^{k-1} + s^{k-1}Ap^{k-1},$$

qui permet de démontrer que

$$\mathrm{Vect}(r^0, r^1, ..., r^k) \subset \mathrm{Vect}(p^0, Ap^0, ..., A^k p^0) = \mathrm{Vect}(r^0, Ar^0, ..., A^k r^0).$$

Or, comme $r^k \perp r^i$ pour tout $i \in \{0, \dots, k-1\}$, la famille des $(r^j)_{0 \leqslant j \leqslant k}$ est libre et donc sa dimension est $k+1$, donc on obtient l'égalité.

- On fait l'hypothèse de récurrence que pour tout $\ell \in \{0, \dots, k-1\}$ et tout $i \in \{0, \dots, \ell - 1\}$, on a $\left\langle p^i, Ap^\ell \right\rangle = 0$. On sait déjà que $\left\langle p^{k-1}, Ap^k \right\rangle = 0$ puisque β^k avait été choisi pour ça. Soit $i \in \{0, \dots, k-2\}$, on calcule

$$\left\langle p^i, Ap^k \right\rangle = -\left\langle p^i, Ar^k \right\rangle + \beta^k \left\langle p^i, Ap^{k-1} \right\rangle = -\left\langle p^i, Ar^k \right\rangle = -\left\langle Ap^i, r^k \right\rangle.$$

On sait que $\left\langle r^k, p^i \right\rangle = 0$ pour tout $i \in \{0, ..., k-1\}$. Or on a aussi

$$Ap^i \in A\,\mathrm{Vect}(r^0, Ar^0, ..., A^i r^0) = \mathrm{Vect}(Ar^0, A^2 r^0, ..., A^{i+1} r^0) \subset \mathrm{Vect}(p^0, p^1, ..., p^{i+1}),$$

donc il existe un vecteur $(\gamma^0, ..., \gamma^{i+1}) \in \mathbb{R}^{i+2}$ tel que $Ap^i = \sum_{j=0}^{i+1} \gamma^j p^j$, et

$$\left\langle Ap^i, r^k \right\rangle = \sum_{j=0}^{i+1} \gamma^j \left\langle p^j, r^k \right\rangle \underset{i+1 \leqslant k-1}{=} 0.$$

□

On peut en outre simplifier les expressions (20) et (25) en utilisant les propriétés de la famille $(r^i)_{0 \leqslant i \leqslant k-1}$. On obtient alors

$$s^k = \frac{\left\|r^k\right\|^2}{\left\langle p^k, Ap^k \right\rangle} \quad \text{et} \quad \beta^{k+1} = \frac{\left\|r^{k+1}\right\|^2}{\left\|r^k\right\|^2} \tag{26}$$

Une fois ces simplifications effectuées on aboutit à l'algorithme suivant

Algorithm 2 Algorithme du gradient conjugué

À partir de $x^0 \in \mathbb{R}^d$ quelconque **calculer** $r^0 = Ax^0 - b$ et $p^0 = -r^0$.
$k \leftarrow 0$
for $i = 1 : d$ **do**
    $s^k = \frac{\left\|r^k\right\|^2}{\left\langle p^k, Ap^k \right\rangle}$
    $x^{k+1} = x^k + s^k p^k$
    $r^{k+1} = r^k + s^k A p^k$
    $\beta^{k+1} = \frac{\left\|r^{k+1}\right\|^2}{\left\|r^k\right\|^2}$
    $p^{k+1} = -r^{k+1} + \beta^{k+1} p^k$
    $k \leftarrow k + 1$
end for
return x

On peut caractériser la vitesse de convergence de cet algorithme. Appelons

$$\kappa := \frac{\lambda_1}{\lambda_d}$$

le rapport entre la plus petite et la plus grande des valeurs propres de A, aussi appelé nombre de conditionnement.

On rappelle qu'on résout le système linéaire $Ax = b$, de solution exact $x_* := A^{-1}b$. On rappelle le produit scalaire $\langle x, y\rangle_A := \langle x, Ay\rangle$.

Proposition 4.9 (Vitesse de convergence du gradient conjugué)**.** *Soit $A \in \mathcal{M}_d(\mathbb{R})$ symétrique définie positive. Notons x_k l'itéré obtenu après k itérations. Alors*

$$\left\|x^k - x_*\right\|_A \leqslant 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^k \left\|x^0 - x_*\right\|_A .$$

Il est possible d'améliorer cette vitesse de convergence par une technique de préconditionnement utilisant un changement de variable linéaire, mais nous ne le ferons pas ici.

4.3.4. *Application aux fonctions non linéaires.* Lorsque la fonction à minimiser n'est pas quadratique, on ne peut pas calculer explicitement le pas optimal s^k le long d'une direction de descente et la notion de résidu n'a pas le même sens. On peut néanmoins transposer les idées de l'algorithme du gradient conjugué avec les algorithmes de Fletcher-Reeves et de Polak-Ribière. On présente ici le premier.

Algorithm 3 Algorithme de Fletcher-Reeves

À partir de $x^0 \in \mathbb{R}^d$ quelconque **calculer** $r^0 = \nabla f(x^0)$ et $p^0 = -r^0$.
$k \leftarrow 0$
while test de convergence non satisfait **do**
    choisir s^k par une recherche linéaire essayant de minimiser $s \mapsto f(x^k + sp^k)$
    $x^{k+1} = x^k + s^k p^k$
    $r^{k+1} = \nabla f(x^{k+1})$ (étape qui change par rapport au gradient conjugué)
    $\beta^{k+1} = \dfrac{\left\|r^{k+1}\right\|^2}{\left\|r^k\right\|^2}$
    $p^{k+1} = -r^{k+1} + \beta^{k+1} p^k$
    $k \leftarrow k + 1$
end while
return x

Partie théorique

5. Convexité

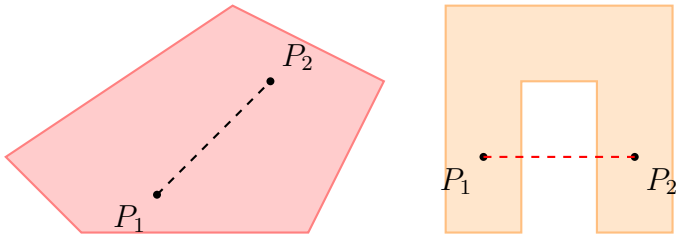
5.1. Ensembles convexes.

Definition 5.1 (Intervalle)**.** *Un intervalle $I \subset \mathbb{R}$ est un sous-ensemble connexe de $\mathbb{R}$ qui n'est pas réduit à un singleton. Il est donc de la forme $[b, +\infty[$ ou $]b, +\infty[$ ou $[a, b]$ ou $[a, b[$ ou $]a, b]$ ou $]a, b[$ ou $]-\infty, a]$ ou $]-\infty, a[$, où $a, b \in \mathbb{R}$ et $a < b$.*

Definition 5.2 (Ensemble convexe)**.** *Un ensemble $K \subset \mathbb{R}^d$ est dit convexe si*

$$\forall x, y \in K,\ \forall t \in [0,1], \qquad tx + (1-t)y \in K.$$

Autrement dit, un ensemble est convexe si quand on prend deux points lui appartenant, le segment reliant ces deux points y est également. En dimension $d = 1$, les ensembles convexes sont les intervalles et les singletons. En figure 2, en dimension $d = 2$, nous dessinons deux exemples.



Ensemble convexe Ensemble non convexe

FIGURE 2. Exemples sur la convexité en dimension 2

5.1.1. *Exemples.*

- $\mathbb{R}^d$ est convexe
- la boule ouverte de centre $a \in \mathbb{R}^d$ et de rayon $R > 0$, $B(a, R) \subset \mathbb{R}^d$, est convexe
- n'importe quel sous-espace affine de $\mathbb{R}^d$ est convexe
- pour $d = 1$, un ensemble $C \subset \mathbb{R}$ est convexe ssi c'est un intervalle
- les demi-espaces affines

$$C := \left\{ x \in \mathbb{R}^d \mid \langle a, x \rangle \leqslant \lambda \right\},$$

où $a \in \mathbb{R}^d \backslash \{0\}$ et $\lambda \in \mathbb{R}$, sont convexes.

5.2. **Fonctions convexes.**

Definition 5.3 (Convexité des fonctions)**.** *Soit $K \subset \mathbb{R}^d$ et une fonction $f : K \to \mathbb{R}$ une fonction. On dit que f est*

- *convexe si*

$$\forall x, y \in K,\ \forall t \in [0,1], \qquad f\left(tx + (1-t)y\right) \leqslant tf(x) + (1-t)f(y).$$

- *strictement convexe si*

$$\forall x, y \in K \text{ tels que } x \neq y,\ \forall t \in]0,1[, \qquad f\left(tx + (1-t)y\right) < tf(x) + (1-t)f(y).$$

- *fortement convexe s'il existe $\alpha > 0$ tel que*

$$\forall x, y \in K,\ \forall t \in [0,1], \quad f\left(tx + (1-t)y\right) \leqslant tf(x) + (1-t)f(y) - \tfrac{1}{2}\alpha t(1-t)\left\|x - y\right\|^2$$

- *concave si $-f$ est convexe.*

On a que fortement convexe $\implies$ strictement convexe $\implies$ convexe. En figure 3 nous donnons quelques exemples. Les fonctions affines sont les seules fonctions convexes et concaves. Visuellement, une fonction est convexe ssi son graph est au-dessus de tous ses plans tangents, ou encore ssi les arcs reliant deux points du graph sont au-dessus du graph.

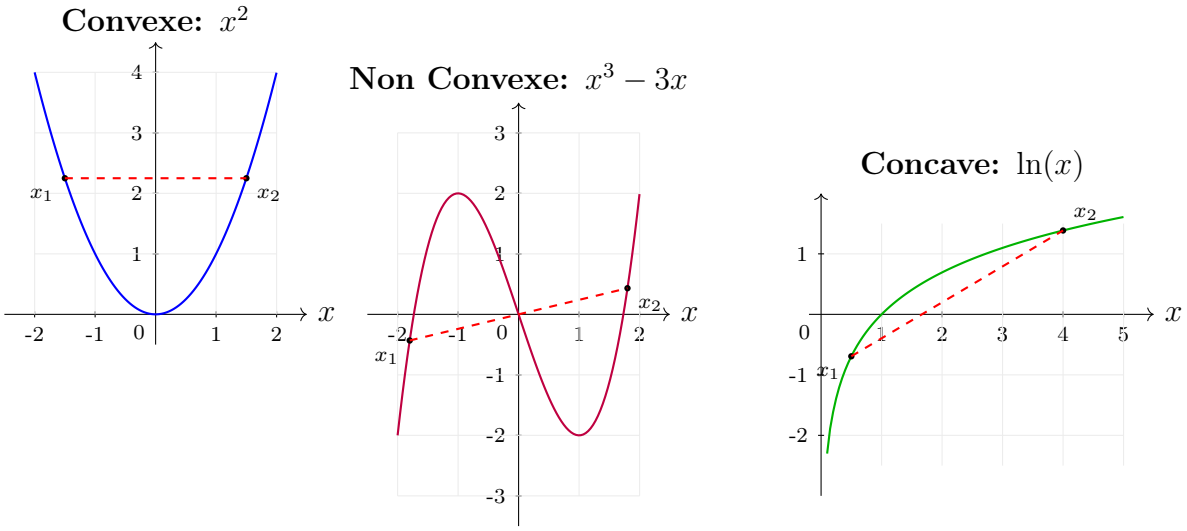


FIGURE 3. Exemples sur la convexité des fonctions en dimension un.

5.2.1. *Fonctions différentiables.* Lorsque f est $\mathcal{C}^1$, c'est-à-dire qu'elle est différentiable et que $x \mapsto \nabla f(x)$ est continue, la convexité peut s'exprimer plus simplement en utilisant le gradient.

Proposition 5.4 (Caractérisation de la convexité avec ∇f)**.** *Soit $K \subset \mathbb{R}^d$ un ensemble convexe et $f : K \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ dans un voisinage de K. Les assertions suivantes sont équivalentes :*

(1) *f est convexe sur K,*
(2) *$f(y) \geqslant f(x) + \langle \nabla f(x), y - x \rangle$, $\forall x, y \in K$,*
(3) *$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geqslant 0$, $\forall x, y \in K$,*
(4) *(Si f est $\mathcal{C}^2$) $\nabla^{\otimes 2} f(x) \geqslant 0$, $\forall x \in K$.*

La seconde assertion signifie que f est au-dessus de son développement d'ordre 1. On peut énoncer le même résultat en remplaçant les inégalités par des inégalités strictes dans les assertions précédentes.

Proof. Prouvons que (1) implique (2). On prend $x, y \in K$. On va se réduire au cas de dimension 1 en posant $\xi(t) := f(x + t(y - x))$. On a $\xi(0) = f(x)$, $\xi(1) = f(y)$. Comme f est convexe, $\xi(t) \leqslant t\xi(1) + (1-t)\xi(0)$, donc la courbe de ξ est en-dessous de la droite passant par $(0, \xi(0))$ et $(1, \xi(1))$. On a donc $\frac{1}{t}(\xi(t) - \xi(0) \leqslant \xi(1) - \xi(0)$ et en faisant $t \to 0$ on obtient $\xi'(0) \leqslant \xi(1) - \xi(0)$. Or, $\xi'(t) = \langle \nabla f(x + t(y - x)), y - x \rangle$ (voir Section (1.4.1)) donc $\xi(0) = \langle \nabla f(x), y - x \rangle$ et on obtient le résultat.

Prouvons que (2) implique (3). On écrit (1) en échangeant x et y puis avec (1) on obtient

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

Prouvons que (2) implique (4). On prend $y = x + tz$, où $t \in \mathbb{R}$, on a

$$f(x + tz) = f(x) + t \langle \nabla f(x), y \rangle + \frac{t^2}{2} \left\langle z, \left(\nabla^{\otimes 2} f(x) \right) z \right\rangle + O(t^3)$$

donc

$$0 \underset{(2)}{\leqslant} \frac{1}{t^2}\left(f(x+tz) - f(x) - t\left\langle \nabla f(x), y\right\rangle\right) = \tfrac{1}{2}\left\langle z, \left(\nabla^{\otimes 2} f(x)\right) z\right\rangle + O(t)$$

on fait $t \to 0$ et on obtient $\langle z, (\nabla^{\otimes 2} f(x))\, z\rangle \geqslant 0$. C'est vrai pour tout $z \in \mathbb{R}^d$ donc on peut conclure. □

La stricte convexité a la même caractérisation en remplaçant les inégalités par des inégalités strictes.

Proposition 5.5 (Caractérisation de la forte convexité avec ∇f)**.** *Soit $K \subset \mathbb{R}^d$ un ensemble convexe et $f : K \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ dans un voisinage de K. Les assertions suivantes sont équivalentes :*

- *f est α-fortement convexe sur K, avec $\alpha > 0$*
- *$f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y - x\|^2, \quad \forall x, y \in K$*
- *$\langle \nabla f(y) - \nabla f(x), y - x\rangle \geqslant \alpha \|y - x\|^2, \quad \forall x, y \in K.$*
- *(Si f est $\mathcal{C}^2$) $\nabla^{\otimes 2} f(x) \geqslant \alpha,\ \forall x \in K$*

La condition $\nabla^{\otimes 2} f(x) \geqslant \alpha$ est équivalente à ce que les valeurs propres de $A^T + A$ soient supérieures à α, ou encore à ce que tous les éléments diagonaux de $D(x)$ (définie en (6)) soient supérieurs à α.

Definition 5.6 (Fonction elliptique)**.** *On appelle fonction elliptique une fonction $f : K \to \mathbb{R}$ de classe $\mathcal{C}^1$ et fortement convexe.*

5.2.2. *Dimension 1.* Les propriétés se simplifient en dimension 1. Soit I un intervalle et $f : I \to \mathbb{R}$ une fonction deux fois différentiable. Alors on a

$$\begin{aligned}
&f \text{ convexe} && \iff && f' \text{ croissante} && \iff && f'' \geqslant 0,\\
&f \text{ strictement convexe} && \iff && f' \text{ strictement croissante} && \iff && f'' > 0,\\
&f \text{ fortement convexe} && \iff && f'' \geqslant \alpha > 0.
\end{aligned}$$

Proposition 5.7 (Propriétés des fonctions convexes en dimension 1)**.** *Soit I un intervalle et $f : I \to \mathbb{R}$ convexe. Alors f est continue et localement lipschitzienne sur $\mathring{I}$. En tout point $a \in \mathring{I}$ (qui signifie l'intérieur de I), f admet une dérivée à gauche $f'_g(a)$ et une dérivée à droite $f'_d(a)$, pas forcément égales. De plus, pour tout $a, b \in \mathring{I}$, on a*

$$f'_g(a) \leqslant f'_d(a) \leqslant \frac{f(b) - f(a)}{b - a} \leqslant f'_g(b) \leqslant f'_d(b).$$

En particulier, les dérivées gauches et droites f'_g et f'_d sont croissantes. Nous ne donnons pas la preuve. La fonction f n'est pas forcément continue sur les bords de I.

5.3. **Exemples.**

Proposition 5.8 (Combinaisons linéaires et compositions)**.**

1. *Soient $f_j : \mathbb{R}^d \to \mathbb{R}$ des fonctions convexes, et $\alpha_j \in \mathbb{R}_+$. Alors $\sum_{j=1}^n \alpha_j f_j$ est convexe.*
2. *Soit $f : \mathbb{R}^d \to \mathbb{R}$ convexe et $g : \mathbb{R} \to \mathbb{R}$ convexe croissante. Alors $g \circ f$ est convexe.*

Preuve de (2). Comme $f(tx+(1-t)y) \leqslant tf(x)+(1-t)f(y)$ pour tout x, $y \in \mathbb{R}$, alors,

$$g\circ f(tx+(1-t)y) \underset{g \text{ croissante}}{\leqslant} g(tf(x)+(1-t)f(y)) \underset{g \text{ convexe}}{\leqslant} tg\circ f(x)+(1-t)g\circ f(y).$$

□

La composée de deux fonctions convexes n'est pas forcément convexe, par exemple $f(x) = -x$ et $g(x) = e^x$ sont convexes mais $g \circ f(x) = e^{-x}$ est concave.

(1) Les normes sont convexes

Proof. Soit $N : \mathbb{R}^d \to \mathbb{R}$ une norme. Pour $t \in [0,1]$ on a

$$N\left(tx+(1-t)y\right) \underset{\substack{\text{ineg.}\\ \text{triang.}}}{\leqslant} N(tx)+N((1-t)y) = tN(x)+(1-t)N(y).$$

□

(2) La fonction $f : \mathbb{R} \to \mathbb{R}$, $f(x) := x^2$ est fortement convexe avec $\alpha = 2$.

Proof. On a pour tout $x \in \mathbb{R}$, $\nabla f(x) = f'(x) = 2x$ et $f''(x) = 2 > 0$. □

(3) La fonction $f : \mathbb{R}_+^* \to \mathbb{R}$, $f(x) := x^p$ est convexe ssi $p \geqslant 1$ ou $p \leqslant 0$.

Proof. On a $f''(x) = p(p-1)x^{p-2}$. Or, $p(p-1) \geqslant 0$ ssi $p \geqslant 1$ ou $p \leqslant 0$. □

(4) La fonction $f : \mathbb{R} \to \mathbb{R}$, $f(x) := e^{\lambda x}$ avec $\lambda > 0$ est strictement convexe mais pas fortement convexe.

Proof. On a $f''(x) = \lambda^2 e^{\lambda x} > 0$ mais $f''(x) \to 0$ quand $x \to -\infty$. □

(5) La fonction $f : \mathbb{R}^d \to \mathbb{R}$, $f(x) := \|x\|^2$ est fortement convexe avec $\alpha = 2$.

Proof. On a $\nabla f(x) = 2x$, on déduit $\langle \nabla f(y) - \nabla f(x), y-x\rangle = 2\|y-x\|^2$. □

(6) Plus généralement, soit f une forme quadratique avec les mêmes notations que la proposition 1.1. On a

$$\langle \nabla f(y) - \nabla f(x), y-x\rangle = \left\langle (A+A^T)(y-x), y-x\right\rangle = 2\left\langle A(y-x),(y-x)\right\rangle.$$

Par conséquent,

(a) A est semi-définie positive $\Leftrightarrow$ f convexe.
(b) A est définie positive de plus petite valeur propre $\lambda_{\min} > 0 \Leftrightarrow$ f fortement convexe avec $\alpha = \lambda_{\min}$.

5.4. **Fonctions coercives.**

Definition 5.9 (Coercivité)**.** *Soit $\Omega \subset \mathbb{R}^d$ un ensemble non borné (par exemple, $\mathbb{R}^d$). Une fonction $f : \Omega \to \mathbb{R}$ est dite coercive sur Ω si*

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

Ceci peut s'écrire de façon équivalente :

- $\forall (x_k)_{k\in\mathbb{N}} \in \Omega^{\mathbb{N}}, \qquad \|x_k\| \to +\infty \implies f(x_k) \to +\infty,$
- $\forall M > 0, \exists R > 0, \forall x \in \Omega \qquad \|x\| \geqslant R \implies f(x) \geqslant M.$

La coercivité ne dépend pas de la norme choisie. Comme toutes les normes sont équivalentes sur $\mathbb{R}^d$, en pratique, on choisit la norme la plus adaptée à la fonction f étudiée.

On verra que la coercivité est importante dans le théorème 6.4.

Proposition 5.10. *Soit $K \subset \mathbb{R}^d$ un ensemble convexe non borné. Si $f : K \to \mathbb{R}$ est $\mathcal{C}^1$ et fortement convexe sur K, alors f est coercive sur K.*

On peut résumer en disant "fortement convexe $\implies$ coercive".

Proof. La seconde assertion de la Proposition 5.5 pour les fonctions fortement convexes nous permet d'écrire pour tout $x, y \in K$,

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{\alpha}{2}\|y - x\|^2. \tag{27}$$

De plus, par l'inégalité de Cauchy-Schwarz,

$$\langle \nabla f(x), x - y\rangle \leqslant \|\nabla f(x)\|\|x - y\|, \quad \text{donc} \quad \langle \nabla f(x), y - x\rangle \geqslant -\|\nabla f(x)\|\|x - y\|.$$

En injectant la dernière inégalité dans (27),

$$f(y) \geqslant f(x) + \|y - x\| \left(\frac{\alpha}{2}\|y - x\| - \|\nabla f(x)\|\right).$$

En fixant $x \in K$ et en faisant $\|y\| \to +\infty$ (ce qui implique $\|y - x\| \to +\infty$), on obtient le résultat. □

Exemples de fonctions coercives:

(1) Par la Proposition 5.10, les formes quadratiques, i.e. les fonctions de la forme
$$f(x) = \langle Ax, x\rangle + \langle b, x\rangle + c, \tag{28}$$
sont fortement convexes, donc coercives, si A est définie positive.

(2) Toute fonction minorée par une fonction coercive est coercive.

Un autre exemple simple nécessite une preuve.

Lemma 5.11. *Soient $f_i : \mathbb{R} \to \mathbb{R}$ des fonctions minorées et coercives. La fonction $f : \mathbb{R}^d \to \mathbb{R}$ donnée par $f(x) = \sum_{i=1}^n f_i(x_i)$ est coercive.*

Proof. Pour tout $i \in \{1, \dots, n\}$, f_i est minorée par une constante m_i. Posons $m = \max_{1\leqslant i\leqslant n} |m_i|$ et soit $M > 0$ fixé. Comme chaque f_i est coercif, il existe une constante $R_i > 0$ telle que

$$\forall x_i \in \mathbb{R}, \quad |x_i| \geqslant R_i \Rightarrow \ f_i(x_i) \geqslant M + nm\ .$$

Soit $R = \max_{1\leqslant i\leqslant n} R_i$. Alors pour tout $x \in \mathbb{R}^d$ avec $\|x\|_\infty \geqslant R$, il existe un indice $i \in \{1, \dots, n\}$ tel que $|x_i| \geqslant R \geqslant R_i$ et donc $f_i(x_i) \geqslant M + nm$. Pour $j \neq i$,

$$f_j(x_j) \geqslant m_j \geqslant -m.$$

On a donc $f(x) \geqslant M$ et on a montré que

$$\forall M \geqslant 0,\ \exists R > 0 \text{ tel que } \forall x \in \mathbb{R}^d,\ \|x\|_\infty \geqslant R \ \Rightarrow \ f(x) \geqslant M\ .$$

Par conséquent, $\lim_{\|x\|_\infty \to \infty} f(x) = +\infty$ ce qui prouve la coercivité de f. □

5.5. **Vocabulaire.** Cherchons les minimas locaux de $f : K \to \mathbb{R}$ où $K \subset \mathbb{R}^d$.

5.5.1. *Sans ou sous contraintes.*

- Si K est un ouvert de $\mathbb{R}^d$, on parle de minimisation sans contrainte pour le problème $\inf_{x\in K} f(x)$. Les extremas sont alors dit libres car ils vérifient l'équation d'Euler (29).
- Sinon, on parle de minimisation sous contrainte. Les extremas sont alors dit liés.

Par exemple si $K = \{x = (x_1, \dots, x_d) \in K \mid x_1 = 0, x_2 \geqslant x_3\}$ on est dans le cadre d'une minimisation sous contraintes.

On parle d'optimisation convexe si f est une fonction convexe et si K est un ensemble convexe.

5.5.2. *Infimum.* Dans cette section on prendra toujours

$$f : K \to \mathbb{R}, \qquad \inf_{x\in K} f(x).$$

On cherche à minimiser f sur K.

Definition 5.12 (Fonction minorée)**.** *On dit que $f : K \to \mathbb{R}$ est minorée sur $K \subset \mathbb{R}^d$ s'il existe $m \in \mathbb{R}$ tel que $\forall x \in K,\ m \leqslant f(x)$. On dit que m est un minorant de f sur K.*

Definition 5.13 (Infimum)**.** *Supposons qu'il existe $\xi \in]-\infty, +\infty[\cup \{-\infty\}$ tel que*

- *$\forall x \in K,\ \xi \leqslant f(x)$*
- *il existe $(x_n)_{n\in\mathbb{N}} \in K^{\mathbb{N}}$ telle que $f(x_n) \underset{n\to+\infty}{\longrightarrow} \xi$.*

Alors ξ est appelé infimum de f sur K et noté $\xi = \inf_{x\in K} f(x) = \inf_K f$.

L'infimum sur K existe toujours, il est fini si et seulement si f est minorée sur K. Si f n'est pas minorée sur K, alors $\inf_K f = -\infty$.

Definition 5.14 (Suite minimisante)**.** *Soit $(x_n)_{n\in\mathbb{N}} \in K^{\mathbb{N}}$ telle que*

$$f(x_n) \underset{n\to+\infty}{\longrightarrow} \inf_K f.$$

Alors $(x_n)_{n\in\mathbb{N}}$ est appelée suite minimisante du problème de minimisation.

5.5.3. *Minimum.*

Definition 5.15 (Minimum global)**.** *S'il existe $x_* \in K$ tel que*

$$\forall x \in K, \qquad f(x_*) \leqslant f(x),$$

alors on dit que $f(x_*)$ est un minimum global de f sur K, on note $f(x_*) = \min_{x\in K} f(x) = \min_K f$. On dit que f atteint son minimum en x_* et que le problème admet x_* comme solution.

Par abus de langage, on dit parfois que x_* est un minimum mais on devrait toujours dire que x_* est un minimiseur. Un minimum existe si et seulement si un minimiseur existe. Le minimum n'existe pas toujours, et on va chercher à trouver des conditions de son existence. S'il existe, le minimum est un infimum.

Definition 5.16 (Minimum local)**.** *Supposons qu'il existe $x_* \in K$ et un voisinage V de x_* dans Ω tels que*

$$\forall x \in V \cap K, \qquad f(x_*) \leqslant f(x).$$

Alors x_* est appelé minimum local de F sur K et $f(x_*)$ est appelé minimum local.

Un minimum global est un minimum local mais le contraire est faux en général. Quand on dit juste "minimum", c'est implicitement qu'on dit minimum global.

6. Les principales conditions d'optimalité

6.1. K ouvert.

Pour tout ensemble $\Omega \subset \mathbb{R}^d$, on appelle $\partial\Omega := \overline{\Omega} \backslash \mathring{\Omega}$ son bord.

Théorème 6.1: Condition suffisante, K ouvert

On suppose que K est un ouvert borné, que f est continue sur $\overline{K}$, et qu'il existe un point $x_0 \in K$ tel que $\forall x \in \partial K,\ f(x_0) < f(x)$. Alors f admet un minimum global sur K.

La condition $\forall x \in \partial K,\ f(x_0) < f(x)$ signifie que si f a un minimum sur $\overline{K}$, il ne sera pas atteint sur le bord, mais dans l'intérieur.

Proof. Comme l'ensemble $\overline{K}$ est compact, la fonction continue f admet un minimum x_* sur $\overline{K}$ par le thèorème de Weierstrass (qu'on reverra au théorème 6.4). Cet élément est donc tel que

$$\forall x \in \overline{K},\ f(x_*) \leqslant f(x).$$

Montrons par l'absurde que $x_* \notin \partial K$. Supposons que $x_* \in \partial K$. On a $f(x_*) \leqslant f(x)$ pour tout $x \in \overline{K}$ donc $f(x_*) \leqslant f(x_0)$. Or, $f(x) > f(x_0)$ pour tout $x \in \partial K$, donc pour $x = x_*$, $f(x_*) > f(x_0)$. Cette contradiction nous permet de conclure que $x_* \notin \partial K$ et donc x_* est bien dans l'ouvert K. □

Exercice 6.1. *Soit $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x) := x_1^2 + x_2^2 + e^{x_2}$. Montrer que f a un minimum sur $B(0,1)$ la boule unité ouverte.*

Solution. On peut paramétrer le bord $\partial B(0,1)$, on a, pour tout $\theta \in [0, 2\pi[$,

$$g(\theta) := f(\cos\theta, \sin\theta) = 1 + e^{\sin\theta} > 1 = f(0).$$

On applique le théorème 6.1 avec $x_0 = 0$. □

Théorème 6.2: Condition nécessaire, K ouvert

Soit $K \subset \mathbb{R}^d$ un ouvert et $f : K \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$. Si x_* est un minimiseur local de f sur K, alors

$$\nabla f(x_*) = 0. \tag{29}$$

Si en plus f est $\mathcal{C}^2$ au voisinage de x_*, alors

$$\nabla^{\otimes 2} f(x_*) \geqslant 0. \tag{30}$$

Proof. Soit y un vecteur quelconque de $\mathbb{R}^d$. Comme $x_* \in K$ et que K est ouvert, il existe $h_0 > 0$ assez petit tel que $B(x_*, h_0) \subset K$, donc le point $x_* + hy$ appartient à K. Or x_* étant un minimum local du problème, on a $0 \leqslant f(x_* + hy) - f(x_*)$. Comme

$$f(x_* + hy) - f(x_*) = h \langle \nabla f(x_*), y \rangle + O(h^2),$$

en divisant l'inegalité ci-dessus par $h > 0$, on obtient $0 \leqslant \langle \nabla f(x_*), y \rangle + O(h)$ donc en faisant $h \to 0$ on a

$$0 \leqslant \langle \nabla f(x_*), y \rangle .$$

Comme cette inegalité est vraie pour tout $y \in \mathbb{R}^d$, elle est également vraie pour $-y$. Donc $\langle \nabla f(x_*), -y \rangle \geqslant 0$, d'où $\langle \nabla f(x_*), y \rangle = 0$. On a montré

$$\langle \nabla f(x_*), y \rangle = 0, \quad \forall y \in \mathbb{R}^d,$$

c'est-à-dire que $\nabla f(x_*) = 0$. Si f est $\mathcal{C}^2$, on a en plus

$$\begin{aligned} f(x_* + hy) - f(x_*) &= \langle \nabla f(x_*), hy \rangle + \tfrac{1}{2} \left\langle hy, \nabla^{\otimes 2} f(x_*) hy \right\rangle + O(h^3) \\ &= \tfrac{1}{2} \left\langle hy, \nabla^{\otimes 2} f(x_*) hy \right\rangle + O(h^3) \end{aligned}$$

On divise par h^2 et on obtient $h^{-2}(f(x_* + hy) - f(x_*)) = \frac{1}{2} \langle y, \nabla^{\otimes 2} f(x_*) y \rangle + O(h)$, or x_* est un minimiseur local donc $f(x_* + hy) - f(x_*) \geqslant 0$ et en faisant tendre $h \to 0$ on obtient la conclusion. □

(29) est appelée équation d'Euler. Ce théorème est très important, il montre que quand on est sur un ouvert, il faut chercher les minimiseurs parmi les points annulant le gradient. Cette procédure donne les minimums locaux, il faudra les comparer entre eux pour trouver le/les minimums globaux.

On notera

$$Z := \{x \in K \mid \nabla f(x) = 0\} = K \cap \left((\nabla f)^{-1} (\{0\}) \right) \tag{31}$$

l'ensemble des zéros du gradient, aussi appelé ensemble des points critiques. En définissant

$$\mathcal{M} := \{x \in \mathbb{R}^2 \mid x \text{ est un minimiseur local}\},$$

le théorème dit que

$$\mathcal{M} \subset Z$$

mais ces ensembles ne sont pas égaux en général. En effet, Z contient

- les minimums locaux
- les maximums locaux
- les points selle

Exercice 6.2. *Trouver les minimiseurs locaux et les minimiseurs globaux de* $f : \mathbb{R}^2 \to \mathbb{R}$,

$$f(x) := x_1^2 + 2x_1 x_2 - x_2^2 + x_2^4.$$

Solution. On cherche $\mathcal{M}$, par le théorème 6.2 on sait que $\mathcal{M} \subset Z$ donc on commence par chercher Z. On calcule

$$\nabla f(x) = 2 \begin{pmatrix} x_1 + x_2 \\ x_1 - x_2 + 2x_2^3 \end{pmatrix}, \qquad \nabla^{\otimes 2} f(x) = 2 \begin{pmatrix} 1 & 1 \\ 1 & 6x_2^2 - 1 \end{pmatrix}.$$

On commence par l'analyse. Si $x \in Z$, alors $x_1 = -x_2$ et $x_1 - x_2 + 2x_2^3$, ce qui implique $x_1 = -x_2$ et $x_2(x_2^2 - 1) = 0$ et donc en définissant $S := \{(0,0), (1,-1), (-1,1)\}$ on a $Z \subset S$. On vérifie facilement que $S \subset Z$ et donc $Z = S$. Trouvons $\mathcal{M}$ par deux méthodes différentes.

- Méthode utilisant la Hessienne. Comme $\mathcal{M} \subset Z$, il faut vérifier pour quels points de Z la Hessienne est définie positive. On a

$$\nabla^{\otimes 2} f(0,0) = 2\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}, \qquad \nabla^{\otimes 2} f(1,-1) = \nabla^{\otimes 2} f(1,-1) = 2\begin{pmatrix} 1 & 1 \\ 1 & 5 \end{pmatrix}.$$

On voit que $\nabla^{\otimes 2} f(0,0)$ a une valeur propre négative donc $(0,0)$ n'est pas un minimiseur local, alors que les valeurs propres de $\nabla^{\otimes 2} f(1,-1)$ sont toutes les deux positives, donc $(1,-1)$ et $(-1,1)$ sont des minimiseurs locaux.

- Méthode sans la Hessienne. On a $f(0,x_2) - f(0,0) = x_2^2(x_2^2 - 1)$ qui est négatif pour x_2 petit, donc $(0,0)$ n'est pas un minimiseur local. On a

$$\begin{aligned} f(x_1,x_2) - f(1,-1) = f(x_1,x_2) - f(-1,1) &= x_1^2 + 2x_1x_2 - x_2^2 + x_2^4 + 1 \\ &\geqslant -2x_2^2 + x_2^4 + 1 \geqslant 0, \end{aligned}$$

donc $(1,-1)$ et $(-1,1)$ sont des minimiseurs locaux.

Dans les deux cas on trouve donc $\mathcal{M} = \{(1,-1),(-1,1)\}$ et on remarque que $\mathcal{M} \neq Z$, $(0,0)$ est en fait un point selle.

On définit

$$\mathcal{G} := \{x \in \mathbb{R}^2 \mid x \text{ est un minimiseur global}\}.$$

On sait que $\mathcal{G} \subset \mathcal{M}$, on a $f(1,-1) = f(1,-1)$. On ne peut pas conclure parce que f pourrait décroître en l'infini. Par l'inégalité de Young avec $p = 3$ et $q = 3/2$, on a $|2x_1x_2| \leqslant \frac{2}{3}\left(|x_1|^{3/2} + 2|x_2|^3\right)$ donc

$$f(x) \geqslant \left(x_1^2 - \tfrac{2}{3}|x_1|^{3/2}\right) + \left(x_2^4 - x_2^2 - \tfrac{4}{3}|x_2|^3\right).$$

On voit que f est coercive et donc $\mathcal{G} = \mathcal{M}$. □

6.1.1. *Condition d'ordre 2.*

Théorème 6.3: Condition suffisante d'ordre 2

On suppose que K est un ouvert et que f est $\mathcal{C}^2$ au voisinage de $x_* \in K$. On suppose aussi que $\nabla f(x_*) = 0$. Si $\nabla^{\otimes 2} f(x_*)$ est définie positive, alors x_* est un miminum local.

Proof. Prenons $y \in \mathbb{R}^d$, on a $\langle y, \nabla^{\otimes 2}(x_*)y\rangle > 0$. Prenons $t > 0$ assez petit pour que $x_* + ty$ reste dans K. Le développement d'ordre 2 donne

$$f(x_* + ty) - f(x_*) = \frac{t^2}{2}\langle y, \nabla^{\otimes 2}(x_*)y\rangle + O(t^3)$$

On divise par t^2 et on obtient $t^{-2}(f(x_* + ty) - f(x_*)) = \frac{1}{2}\langle y, \nabla^{\otimes 2}(x_*)y\rangle + O(t)$, donc pour t assez petit, on a $t^{-2}(f(x_* + ty) - f(x_*)) \geqslant 0$. □

Si on a une direction $y \in \mathbb{R}^d$ dans laquelle la Hessienne est dégénérée, i.e. si $\langle y, \nabla^{\otimes 2} f(x_*) y\rangle = 0$, alors $f(x_* + y) = f(x_*) + O\left(\|y\|^3\right)$. L'ordre 2 ne suffit alors pas pour connaître le signe de $f(x_* + y) - f(x_*)$. Il faut donc regarder l'ordre 3.

Pour un point critique $x_* \in K$,

- Si $\nabla^{\otimes 2} f(x_*)$ a toutes ses valeurs propres strictement positives alors x_* est un minimum local
- Si $\nabla^{\otimes 2} f(x_*)$ a toutes ses valeurs propres strictement négatives alors x_* est un maximum local
- Si $\nabla^{\otimes 2} f(x_*)$ a au moins une valeur propre nulle, on ne peut pas conclure pour l'existence d'un maximum ou minimum local
- Si $\nabla^{\otimes 2} f(x_*)$ a au moins une valeur propre positive et au moins une négative, alors x_* est un point selle

6.2. K fermé.

Théorème 6.4: Condition suffisante, K fermé

Soit $K \subset \mathbb{R}^d$ un ensemble non-vide et fermé dans $\mathbb{R}^d$ et $f : K \to \mathbb{R}$ une fonction continue. Si K est borné ou f est coercive sur K, alors f admet un minimum global sur K.

C'est pour ce théorème que la coercivité est une propriété importante des fonctions. Rappelons que dans le cas où K est borné, f a aussi un maximum sur K.

Proof. Dans le premier cas, K est fermé et borné, donc il est compact. Comme f est continue, le théorème de Weierstrass assure que f est bornée sur K et elle atteint ses bornes. Donc il existe au moins un minimiseur.

Prouvons le second cas. Soit $x_0 \in K$ fixé. La coercivité de f sur K entraîne qu'il existe $r > 0$ tel que

$$\forall x \in K, \quad \|x\| \geqslant r \quad \Rightarrow f(x) > f(x_0). \tag{32}$$

On note $\overline{B(0,r)}$ la boule fermée de $\mathbb{R}^d$ de centre 0 et de rayon r. Sans perte de généralité, on prend r assez grand pour que $x_0 \in \overline{B(0,r)}$. Nous avons

$$\inf_{x \in K} f(x) = \inf_{x \in K \cap \overline{B(0,r)}} f(x).$$

Comme l'ensemble $K \cap \overline{B(0,r)}$ est fermé et borné et que f est continue, le théorème de Weierstrass assure que f atteint ses bornes dans $K \cap \overline{B(0,r)}$. Ceci assure l'existence d'un minimum x_* dans $K \cap \overline{B(0,r)}$. Ce minimum est aussi le minimum sur K. En effet, pour tout $x \in K$:

(1) soit $x \in K \cap \overline{B(0,r)}$, et alors $f(x) \geqslant f(x_*)$ car f atteint son minimum sur $K \cap \overline{B(0,r)}$ en x_*,

(2) soit $x \notin K \cap \overline{B(0,r)}$, auquel cas on a $f(x) > f(x_0) \geqslant f(x_*)$ où la seconde inégalité vient du fait que x_0 est dans la boule.

Ceci montre donc que x_* est un minimum de f sur K. □

Exercice 6.3. *Soit $f : \mathbb{R}^3 \to \mathbb{R}$, $f(x) := -x_1 x_2^2 + |x_1|^3 + e^{x_2} + |x_3|$. Montrer que f admet un minimum sur $\mathbb{R}^3$.*

Solution. On a

$$f(x) \geqslant -2(x_1^2 + x_2^4) + |x_1|^3 + e^{x_2} + |x_3| = x_1^2(|x_1| - 2) + (e^{x_2} - 2x_2^4) + |x_3|\,,$$

donc f est coercive, et on peut conclure en utilisant le théorème 6.4. □

En pratique, sur un fermé borné K, pour trouver les minimums, il faudra chercher en deux étapes :

- dans l'intérieur $\mathring{K}$ qui est un ouvert. On cherchera donc dans ce cas les points qui annulent le gradient, on note $Z := \{x \in \mathring{K} \mid \nabla f(x) = 0\}$.
- sur le bord ∂K.

Autrement dit,

$$\min_{x \in K} f(x) = \min_{x \in Z \cap \partial K} f(x).$$

6.3. K **convexe.** Quand K est convexe, on a des conditions nécessaires et suffisantes utilisant le gradient.

Théorème 6.5: K convexe

Soit K un ensemble convexe et $f : K \to \mathbb{R}$ une fonction $\mathcal{C}^1$. Si x_* est un minimiseur local de f sur K, alors

$$\forall y \in K, \qquad \langle \nabla f(x_*), y - x_* \rangle \geqslant 0. \tag{33}$$

Si, en plus f est convexe, alors,

$$(33) \iff \quad x_* \text{ est un minimiseur global de } f \text{ sur } K.$$

L'inéquation (33) est appelée inéquation d'Euler. Elle signifie que partant de x_*, f croît dans toutes les directions qui restent dans K.

Avant de présenter ce résultat, il est important de voir que la condition (33) se réduit à l'équation d'Euler

$$\nabla f(x_*) = 0.$$

lorsque K est un ouvert (en particulier, lorsque $K = \mathbb{R}^d$).

Proof.

- Prouvons la première partie. Soit $y \in K$ et $h \in]0, 1]$. Alors, $x_* + h(y - x_*) \in K$ car K est convexe. De plus, comme x_* est un minimum local de K,

$$\frac{f(x_* + h(y - x_*)) - f(x_*)}{h} \geqslant 0.$$

Or on a le développement de Taylor à l'ordre 1

$$\begin{aligned} f(x_* + h(y - x_*)) &= f(x_*) + (\mathrm{d}_{x_*} f)\,(h(y - x_*)) + O(h^2) \\ &= f(x_*) + h\,\langle \nabla f(x_*), y - x_* \rangle + O(h^2), \end{aligned}$$

donc $\langle \nabla f(x_*), y - x_* \rangle + O(h) \geqslant 0$. Enfin, on fait $h \to 0$.

- Prouvons la seconde partie. Le sens $\Longleftarrow$ se fait avec la partie précédente. Montrons le sens $\Longrightarrow$. Comme f est convexe sur K, pour tout $y \in K$,

$$f(y) \geqslant f(x_*) + \langle \nabla f(x_*), y - x_* \rangle$$

grâce à la formule (ii) de la Proposition 5.4. Comme $\langle \nabla f(x_*), y - x_* \rangle \geqslant 0$, on a donc $f(y) \geqslant f(x_*)$ pour tout $y \in K$ ce qui prouve que x_* est un minimum global de f sur K. □

Dans le cas où K est convexe et f est non seulement convexe, mais strictement ou fortement convexe, nous avons les deux résultats suivants qui permettront de garantir existence, unicité et caractérisation du minimiseur.

Théorème 6.6: Existence et unicité, K convexe

Soit $K \subset \mathbb{R}^d$ un ensemble convexe et $f : K \to \mathbb{R}^d$ une fonction strictement convexe. S'il existe un minimiseur sur K, alors il est unique.
Si en plus K est fermé et si f est $\mathcal{C}^1$ et fortement convexe, alors il existe un unique minimiseur global.

Proof.

- Nous allons raisonner par l'absurde. Soient x_1 et $x_2 \in K$ avec $x_1 \neq x_2$ deux points de minimum de f sur K. Nous avons donc $f(x_1) = f(x_2) \leqslant f(x)$ pour tout $x \in K$. Comme f est strictement convexe et que $(x_1+x_2)/2 \in K$ car K est convexe, nous avons

$$f\left(\tfrac{x_1+x_2}{2}\right) < \tfrac{1}{2}f(x_1) + \tfrac{1}{2}f(x_2) = f(x_1)$$

Ceci contredit le fait que x_1 soit un minimum.

- Comme f est fortement convexe et de classe $\mathcal{C}^1$, elle est continue et coercive. L'existence du minimum est garantie par le théorème 6.4. □

Par le théorème 6.5, ce minimiseur vérifie l'inéquation d'Euler (33).

Exercice 6.4. *On définit $f(x_1, x_2) := (x_1 + x_2)^2 + x_1x_2 - 4x_1 - 3x_2$, montrer que f a un unique minimum sur $K := \{x \in \mathbb{R}^2 \mid x_1 + x_2 \leqslant 3\}$.*

Solution. K et f sont convexes, K est fermé, f est une forme quadratique avec $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ qui a ses valeurs propres strictement positives donc f est fortement convexe donc on peut appliquer le théorème 6.6. □

7. Optimisation sous contraintes d'égalités et inégalités

Nous commencerons par énoncer le théorème KKT. Ensuite nous le démontrerons dans quelques cas simples, et enfin nous donnerons un exemple d'application.

7.1. Théorème KKT.

Definition 7.1 (Contraintes d'égalités et d'inégalités)**.** *Définissions des fonctions $g_i \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ (pour $i \in \{1, \dots, m\}$) et $h_i \in \mathcal{C}^1(\mathbb{R}^d, \mathbb{R})$ (pour $i \in \{1, \dots, p\}$), et*

$$G : \begin{array}{ccc} \mathbb{R}^d & \longrightarrow & \mathbb{R}^m \\ x & \longmapsto & G(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{pmatrix} \end{array} \quad \Bigg| \quad H : \begin{array}{ccc} \mathbb{R}^d & \longrightarrow & \mathbb{R}^p \\ x & \longmapsto & G(x) = \begin{pmatrix} h_1(x) \\ \vdots \\ h_p(x) \end{pmatrix} \end{array}$$

Le domaine

$$K := \{x \in \mathbb{R}^d,\ G(x) \leqslant 0,\ H(x) = 0\}, \tag{34}$$

est défini par contraintes d'égalités (via H) et d'inégalités (via G). L'inégalité $G(x) \leqslant 0$ signifie que $g_i(x) \leqslant 0$ pour tout $i \in \{1, \dots, m\}$.

Le domaine K est alors fermé mais pas borné ni convexe en général.

Le théorème KKT (Karush, Kuhn et Tucker) donne des conditions nécessaires d'optimalité lorsque la contrainte K est de la forme (34).

Nous notons

$$I(x) := \{i \in \{1, \dots, m\} \mid g_i(x) = 0\} \subset \{1, \dots, m\}$$

l'ensemble des contraintes d'inégalité actives au point x.

Definition 7.2 (Qualification des contraintes)**.** *Les contraintes* (34) *sont qualifiées au point $x \in K$ si toutes les contraintes sont affines, ou si les deux conditions suivantes sont satisfaites*

(1) *les vecteurs $(\nabla h_j(x))_{j=1}^p$ sont linéairement indépendants*
(2) *il existe une direction $\omega \in \mathbb{R}^n$ telle que*

$$\langle \nabla h_j(x), \omega \rangle = 0, \quad \forall j \in \{1, \dots, p\}$$
$$\langle \nabla g_i(x), \omega \rangle < 0, \quad \forall i \in I(x).$$

La condition $\langle \nabla h_j(x), \omega \rangle = 0$ signifie que $h_i(x)$ ne change pas dans la direction ω, et donc reste à 0, ainsi, ceci assure que localement, $H(x) = 0$ est conservé. La condition $\langle \nabla g_i(x), \omega \rangle < 0$ signifie que g décroît strictement dans la direction ω, ceci assure que $G(x) \leqslant 0$ est conservé.

Nous pouvons maintenant énoncer le théorème KKT.

Théorème 7.1: Théorème KKT

Soit K comment en (34) et $x_* \in K$, et on considère $f : \mathbb{R}^d \to \mathbb{R}$. On suppose que f, G et H sont $\mathcal{C}^1$ et que les contraintes sont qualifiées en x_* au sens de la définition 7.2. Si x_* est un minimiseur local de f sur K, alors il existe $\mu_1, \dots, \mu_p \in \mathbb{R}$ et $\lambda_1, \dots, \lambda_m \in \mathbb{R}$ tels que pour tout $i \in \{1, \dots, m\}$,

$$\nabla f(x_*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x_*) + \sum_{j=1}^{p} \mu_j \nabla h_j(x_*) = 0, \qquad \lambda_i \geqslant 0, \qquad \lambda_i g_i(x_*) = 0. \tag{35}$$

L'équation (35) est une condition nécessaire à l'optimalité. Introduisons le lagrangien

$$\begin{array}{rccl} \mathcal{L} : & \mathbb{R}^d \times (\mathbb{R}_+)^m \times \mathbb{R}^p & \longrightarrow & \mathbb{R} \\ & (x, \lambda, \mu) & \longmapsto & f(x) + \sum_{i=1}^m \lambda_i g_i(x) + \sum_{j=1}^p \mu_j h_j(x). \end{array}$$

Les réels μ_i et λ_i de (35) sont appelés multiplicateurs de Lagrange, ce sont les facteurs des contraintes. Les conditions (35) peuvent se réécrire

$$(\nabla_x \mathcal{L})(x_*, \lambda, \mu) = 0, \qquad \lambda \geqslant 0, \qquad \lambda \cdot G(x_*) = 0.$$

Essayons de comprendre ce résultat.

7.2. **Introduction.** Commençons par définir l'ensemble des directions admissibles. C'est l'ensemble des directions telles que si on les suit en partant de x, alors on reste dans K au premier ordre.

Definition 7.3 (Directions admissibles)**.** *Soit $x \in K$. On définit l'ensemble des directions de déplacement admissibles au point x comme étant*

$$D_K(x) := \left\{\omega \in \mathbb{R}^n \mid \exists\, 0 < t_k \to 0^+,\, \exists\, \omega_k \to \omega,\ \text{tq } \forall k \in \mathbb{N},\ x + t_k\omega_k \in K\right\}.$$

Un cône est un ensemble E tel que $x \in E \implies (\forall \lambda \geqslant 0, \lambda x \in E)$. On peut facilement montrer que $D_K(x)$ est un cône.

On rappelle le principe suivant de minimisation.

Lemma 7.4 (Principe général d'optimisation)**.** *Si $x_* \in K$ est un minimiseur local de f, alors aucune direction de déplacement admissible n'est une direction de diminution stricte de f. En termes mathématiques,*

$$y \in D_K(x_*) \implies \langle \nabla f(x_*), y\rangle \geqslant 0. \tag{36}$$

Il va falloir transformer (36) en (35).

7.3. **Contraintes d'égalité.** Pour comprendre le théorème KKT, regardons d'abord ce qu'il se passe quand il n'y a que des contraintes d'égalité, c'est-à-dire quand

$$K = \{v \in \mathbb{R}^d \mid H(v) = 0\}. \tag{37}$$

7.3.1. *Cas où K est un espace affine.* Supposons que K est un espace affine de $\mathbb{R}^d$, c'est donc un ensemble convexe fermé non borné. On explicite la forme de K. On peut exprimer $K = x_0 + V = \{x_0 + v, v \in V\}$ avec V un sous-espace vectoriel de $\mathbb{R}^d$ et $x_0 \in K$, x_0 peut être choisi comme étant n'importe quel point de K. On peut toujours exprimer un espace vectoriel de dimension finie comme une intersection d'hyperplans, donc il existe $p \leqslant n$ et $u_1, \dots, u_p \in \mathbb{R}^d$ tels que

$$V = \{x \in \mathbb{R}^d \mid \langle u_i, x\rangle = 0,\ 1 \leqslant i \leqslant p\} = \left(\operatorname{Span}(u_i)_{i=1}^p\right)^\perp.$$

Si $f : K \to \mathbb{R}$ est $\mathcal{C}^1$, l'inéquation d'Euler (33) garantit que si x_* est un minimiseur, alors $\langle \nabla f(x_*), y - x_*\rangle \geqslant 0$ pour tout $y \in K$. On a $K = x_* + V$ donc $\{y - x_*, y \in K\} = V$. La condition d'Euler donne alors $\langle \nabla f(x_*), v\rangle \geqslant 0$ pour tout $v \in V$. En échangeant v par $-v \in V$, on obtient $\langle \nabla f(x_*), v\rangle = 0$ pour tout $v \in V$, ce qui équivaut à $\nabla f(x_*) \in V^\perp = \left(\left(\operatorname{Span}(u_i)_{i=1}^p\right)^\perp\right)^\perp = \operatorname{Span}(u_i)_{i=1}^p$, i.e.

$$\exists\, \mu_1, \dots, \mu_p \in \mathbb{R} \text{ tels que } \nabla f(x_*) + \sum_{i=1}^p \mu_i u_i = 0.$$

On voit de la manière la plus simple possible l'apparition des multiplicateurs de Lagrange.

7.3.2. *Contraintes d'égalités générales.* Localement, tout bord d'un fermé est un espace affine, donc on va se ramener au cas précédent.

Considérons le cas général (37). Il n'est pas possible d'utiliser l'inéquation d'Euler puisque K n'est pas forcément convexe. Cependant, il est possible de généraliser en cherchant quelles sont les directions admissibles pour lesquelles l'inéquation reste vraie. On a

$$D_K(x) = \mathrm{T}_x K = \mathrm{Ker}\, \mathrm{d}_x H = \{v \in \mathbb{R}^d \mid \langle \nabla h_i(x), v\rangle = 0,\ 1 \leqslant i \leqslant p\} \tag{38}$$

$$= (\mathrm{Span}(\nabla h_i(x_*))_{i=1}^p)^\perp, \tag{39}$$

où $\mathrm{T}_x K$ est l'espace tangent de K en x. On admettra le résultat suivant.

Proposition 7.5. *Soit x_* un minimum local de f sur l'ensemble K défini en* (37). *Si la famille de vecteurs $(\nabla h_i(x_*))_{i=1}^p$ est libre, alors*

$$\langle \nabla f(x_*), y\rangle = 0, \quad \forall y \in D_K(x_*).$$

Dans le cas présent, $D_K(x_*)$ est plus qu'un cône, il s'agit du plan tangent à la variété K au point x_*. La proposition précédente nous permet d'énoncer le résultat suivant.

Theorem 7.6. *Soit K comme en* (37) *et $x_* \in K$. Supposons que les fonctions h_i soient $\mathcal{C}^1$ dans un voisinage de x_* pour tout $i \in \{1, \dots, p\}$. De plus, supposons que les vecteurs $(\nabla h_i(x_*))_{i=1}^p$ soient linéairement indépendants. Si x_* est un minimiseur local de f sur K, alors il existe $\mu_1, \dots \mu_p \in \mathbb{R}$ tels que*

$$\nabla f(x_*) + \sum_{i=1}^p \mu_i \nabla h_i(x_*) = 0. \tag{40}$$

Proof. Les hypothèses de la proposition 7.5 étant satisfaites, alors $\langle \nabla f(x_*), y\rangle = 0$ pour tout $y \in D_K(x_*)$. Donc $\nabla f(x_*) \in \Big((\mathrm{Span}(\nabla h_i(x_*))_{i=1}^p)^\perp\Big)^\perp = \mathrm{Span}(\nabla h_i(x_*))_{i=1}^p$. □

7.4. **Contraintes d'inégalité.** On va traiter maintenant le cas d'inégalité pur.

$$K = \{v \in \mathbb{R}^d \mid G(v) \leqslant 0\}. \tag{41}$$

7.4.1. *Cas où K est un cône.* Supposons que

$$K = \{x \in \mathbb{R}^d \mid \langle v_i, x\rangle \leqslant 0,\ 1 \leqslant i \leqslant m\}$$

où $v_i \in \mathbb{R}^d$. En appliquant l'inéquation d'Euler (33) à $y = 0$ et $y = 2x_*$, on a $\langle \nabla f(x_*), x_*\rangle = 0$, et la condition d'Euler (33) devient $\langle \nabla f(x_*), y\rangle = 0, \forall y \in K$. Il est possible de prouver par le lemme de Farkas (que l'on ne présentera pas) que les conditions nécessaires d'optimalité sont

$$\exists \lambda_1, \dots, \lambda_m \geqslant 0 \text{ tels que} \qquad \nabla f(x_*) + \sum_{i=1}^m \lambda_i v_i = 0. \tag{42}$$

En prenant le produit scalaire de (42) avec x_*, et en utilisant $\langle \nabla f(x_*), x_*\rangle = 0$, on voit que si $\langle v_i, x_*\rangle < 0$, alors $\lambda_i = 0$. Les $\lambda_i \geqslant 0$ sont de nouveau appelés multiplicateurs de Lagrange.

7.4.2. *Cas d'une seule contrainte.* Commençons par étudier le cas où il y a une seule contrainte d'inégalité,

$$K = \{v \in \mathbb{R}^d \mid g(v) \leqslant 0\}. \tag{43}$$

Lemma 7.7 (Directions admissibles en cas de contraintes d'inégalités)**.** *Soit $x \in K$. Dans le cas* (43)*, on a*

$$D_K(x) \begin{cases} = \mathbb{R}^d & \text{si } g(x) < 0, \\ \supset \{\omega \in \mathbb{R}^d \mid \langle \nabla g(x), \omega \rangle < 0\} & \text{si } g(x) = 0. \end{cases}$$

Proposition 7.8. *Soit x_* un minimiseur local de f sur K. Si $\nabla g(x_*) \neq 0$ alors il existe $\lambda \geqslant 0$ tel que $\lambda g(x_*) = 0$ et $\nabla f(x_*) + \lambda \nabla g(x_*) = 0$.*

Proof. Supposons que $g(x_*) < 0$. Par le Lemme 7.4, il n'y a aucune direction $\omega \in \mathbb{R}^d$ telle que $\langle \nabla f(x_*), \omega \rangle < 0$, et par le Lemme 7.7 toutes les directions sont admissibles donc $\nabla f(x_*) = 0$. Les deux conclusions sont vérifiées pour $\lambda = 0$.

Supposons que $g(x_*) = 0$. Par les Lemmes 7.4 et 7.7 il n'y a aucune direction $\omega \in \mathbb{R}^d$ telle qu'on ait à la fois $\langle \nabla g(x_*), \omega \rangle < 0$ (direction admissible) et $\langle \nabla f(x_*), \omega \rangle < 0$ (direction de diminution de f). Si $\nabla g(x_*) \neq 0$, alors il existe $\lambda \geqslant 0$ tel que $\nabla f(x_*) = -\lambda \nabla g(x_*)$. Sinon, il existerait un vecteur ω formant un angle obtus à la fois avec $\nabla f(x_*)$ et $\nabla g(x_*)$. Un petit déplacement dans la direction ω permettrait alors de diminuer strictement f tout en restant dans K. Les deux équations conclusion sont vérifiées. □

7.4.3. *K avec contraintes d'inégalité.* Prenons maintenant le cas (41). De façon similaire au cas avec contraintes d'égalité, la démarche consiste à trouver pour chaque point x le cône de directions admissibles $D_K(x)$ où il est possible de formuler une équation ou inéquation d'Euler. Dans les cas précédents, les gradients des contraintes ont suffi pour identifier $D_K(x)$. Afin que cela reste vrai dans le cas présent, il est nécessaire d'ajouter des conditions supplémentaires sur les contraintes, qui sont les conditions de qualification de la définition 7.2. Ceci provient du fait que dans le cas actuel, toutes les contraintes ne vont pas jouer le même rôle en chaque point x et il est nécessaire d'ajouter des conditions supplémentaires dans le but de garantir que les gradients des contraintes déterminent entièrement les directions dans lesquelles il est possible de faire des variations autour d'un point. Dit d'une façon plus abstraite, les conditions de qualification garantissent que $D_K(x)$ peut s'obtenir en linéarisant les contraintes.

Definition 7.9 (Qualification dans le cas d'inégalités)**.** *Les contraintes* (41) *sont qualifiées au point $x \in K$ si $I(x) = \varnothing$ ou bien s'il existe une direction $\omega \in \mathbb{R}^d$ telle que*

$$\begin{aligned} &\text{soit } \langle \nabla g_i(x), \omega \rangle < 0, \quad \forall i \in I(x). \\ &\text{soit } \langle \nabla g_i(x), \omega \rangle = 0 \text{ et } g_i \text{ est affine.} \end{aligned} \tag{44}$$

Remark 7.10. *Les contraintes sont automatiquement qualifiées dans deux cas importants, ce qui facilite grandement l'analyse de nombreux cas pratiques:*

- *Si toutes les fonctions g_i sont affines, nous pouvons prendre $\omega = 0$, ce qui satisfait automatiquement la condition de qualification.*

- *Plus généralement, si toutes le contraintes g_i sont convexes et $\mathcal{C}^1$, la contrainte K est automatiquement qualifiée si $\mathring{K} \neq \varnothing$.*

  Proof. Fixons $x_0 \in \mathring{K}$. Soit $x \in \partial K$ et posons $\omega := x_0 - x \neq 0$. Comme la contrainte g_i est convexe, on a pour tout $i \in I(x)$,
  $$\langle \nabla g_i(x), \omega \rangle \leqslant g_i(x_0) - g_i(x) = g_i(x_0) < 0.$$

□

Theorem 7.11. *Soit K comme dans (41) où les g_i sont $\mathcal{C}^1$, et $x_* \in K$. Supposons que les contraintes sont qualifiées en x_*. Si x_* est un minimum local de f sur K, alors il existe $\lambda_1, \dots, \lambda_m \geqslant 0$ tels que $\forall i \in \{1, \dots, m\}$,*
$$\nabla f(x_*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x_*) = 0, \qquad \lambda_i \geqslant 0, \qquad \lambda_i = 0 \text{ si } g_i(x_*) < 0. \tag{45}$$

Remark 7.12. *La condition*
$$\lambda_i \geqslant 0, \qquad \lambda_i = 0 \text{ si } g_i(x_*) < 0, \qquad \forall i \in \{1, \dots, m\}$$
peut s'écrire de façon équivalente comme
$$\lambda_i \geqslant 0, \quad \lambda_i g_i(x_*) = 0, \qquad \forall i \in \{1, \dots, m\}$$
ou encore
$$\lambda \geqslant 0, \quad \lambda \cdot G(x_*) = 0.$$
La condition $\lambda \cdot G(x_*) = 0$, est appelée condition d'exclusion.

7.5. **Le cas convexe.** Dans cette section, nous nous concentrons sur le cas convexe, au sens de la définition suivante.

Definition 7.13 (Problème convexe)**.** *On dit que le problème de minimisation $\min_{x \in K} f(x)$ est convexe si f et les g_i sont convexes et $\mathcal{C}^1$, et si les h_i sont affines.*

Nous allons prouver que dans ce cas, les conditions KKT données en (35) sont non seulement nécessaires, mais aussi suffisantes, et elles sont suffisantes sans aucune condition de qualification. La qualification reste malgré tout nécessaire pour énoncer la condition nécessaire, mais nous allons voir que dans ce cas elle s'exprime de façon très simple.

Theorem 7.14 (CNS Cas Convexe)**.** *Dans le cas convexe,*
$$\begin{cases} \nabla f(x_*) + \sum_{i=1}^{m} \lambda_i \nabla g_i(x_*) + \sum_{j=1}^{p} \mu_j \nabla h_j(x_*) = 0, \\ \lambda_i \geqslant 0, \quad \lambda_i g_i(x_*) = 0, \ \forall i = 1, \dots, m. \end{cases}$$
$$\overset{\Leftarrow \text{ qualif.}}{\underset{\text{sans qualif.} \Rightarrow}{\Longleftrightarrow}} \quad x_* \text{ solution}$$

Proof. L'implication réciproque étant garantie par le Théorème 7.1, il suffit de montrer l'implication directe. Pour cela, soit $x_* \in K$ un point satisfaisant la relation KKT. On a donc
$$\nabla f(x_*) = -\sum_{i=1}^{m} \lambda_i \nabla g_i(x_*) - \sum_{j=1}^{p} \mu_j \nabla h_j(x_*). \tag{46}$$

La fonction f étant convexe sur K,

$$f(x) - f(x_*) \geqslant \langle \nabla f(x_*), x - x_* \rangle , \quad \forall x \in K$$

et en remplaçant $\nabla f(x_*)$ par la formule (46), on a

$$f(x) - f(x_*) \geqslant \sum_{i=1}^{m} \lambda_i \langle \nabla g_i(x_*), x_* - x \rangle + \sum_{i=1}^{p} \mu_i \langle \nabla h_i(x_*), x_* - x \rangle . \tag{47}$$

Les fonctions h_i étant affines, elles peuvent s'exprimer comme

$$h_i(x) = h_i(x_*) + \langle \nabla h_i(x_*), x - x_* \rangle .$$

Or, comme x et x_* sont dans K, on a $h_i(x) = h_i(x_*) = 0$ car ce sont des contraintes d'égalité. Donc $\langle \nabla h_i(x_*), x - x_* \rangle = 0$ et l'inégalité (47) devient

$$f(x) - f(x_*) \geqslant \sum_{i=1}^{m} \lambda_i \langle \nabla g_i(x_*), x_* - x \rangle \geqslant \sum_{i=1}^{m} \lambda_i (g_i(x_*) - g_i(x)),$$

où nous avons utilisé le fait que les contraintes g_i sont convexes. Par ailleurs, comme chaque $\lambda_i \geqslant 0$ et $g_i(x) \leqslant 0$ pour tout $x \in K$, on a donc $\sum_{i=1}^{m} \lambda_i g_i(x) \leqslant 0$ et donc

$$f(x) - f(x_*) \geqslant \sum_{i=1}^{m} \lambda_i g_i(x_*).$$

Finalement, par la condition d'exclusion, si $g_i(x_*) < 0$ alors $\lambda_i = 0$ et sinon $g_i(x_*) = 0$, d'où la conclusion $f(x) \geqslant f(x_*)$. □

Les conditions de qualification dans le cas convexe sont en général très simples à examiner comme le montre la proposition suivante.

Proposition 7.15. *Supposons que $\mathring{K} \neq \varnothing$. Si les vecteurs $(\nabla h_j(x))_{j=1}^{p}$ sont linéairement indépendants, alors les contraintes sont qualifiées au point $x \in K$.*

Proof. Soit $x \in K$ fixé et soit x_0 un point de l'intérieur de K. En vue de la Définition 7.2 sur la qualification, il suffit de vérifier qu'il existe une direction $\omega \in \mathbb{R}^d$ telle que

$$\langle \nabla h_j(x), \omega \rangle = 0, \quad \forall j \in \{1, \dots, p\} \qquad \text{et} \qquad \langle \nabla g_i(x), \omega \rangle < 0, \quad \forall i \in I(x).$$

Prenons $\omega = x_0 - x$. La remarque 7.10 permet d'affirmer que la condition sur les g_i est vérifiée. De plus, suivant un raisonnement similaire à celui donné dans la preuve précédente, nous avons $\langle \nabla h_j(x), \omega \rangle = h(x_0) - h(x) = 0$ car les h_i sont affines et que $h(x) = h(x_0) = 0$ vu que x, $x_0 \in K$. □

7.6. **Pour résoudre un problème de minimisation.** Il y a 4 étapes

(1) On montre *a priori* que le problème admet une solution, en utilisant les résultats des sections précédentes.
(2) On détermine l'ensemble E_1 des points qui ne vérifient pas la définition 7.2, i.e. n'étant pas qualifés.
(3) Parmi les points qualifiés, on cherche les points satisfaisant les conditions nécessaires de KKT (35). On note cet ensemble E_2.
(4) Si le problème a une solution, le minimum appartient à $E_1 \cup E_2$, et on le cherche dans cet ensemble.

Dans le cas convexe, il suffit de trouver des points vérifiant les conditions nécessaires d'optimalité pour pouvoir conclure à l'existence d'un minimum (voir théorème 7.14).

7.7. **Exemple.** On définit

$$f(x,y) := x - 2y, \qquad K := \{(x,y) \in \mathbb{R}^2,\ -x^2 \leqslant y \leqslant 0,\ -1 \leqslant x \leqslant 0\},$$

et on considère le problème $\min_{x \in K} f(x)$. Cet exemple ne fait intervenir que des contraintes d'inégalité.

7.7.1. *Énoncé du problème.*

(1) Montrer que le problème admet au moins une solution.
(2) Dessiner la contrainte K.
(3) Montrer que, pour tout point $(x,y) \in K$ avec $(x,y) \neq (0,0)$, la contrainte est qualifiée en (x,y).
(4) Trouver l'ensemble des points vérifiant les conditions nécessaires d'optimalité.
(5) Déterminer la ou les solutions du problème.

7.7.2. *Solution.* **1.** La contrainte K est fermée car définie par des inégalités larges faisant intervenir des fonctions polynomiales continues. Elle est aussi bornée puisque si $(x,y) \in K$, alors $|x| \leqslant 1$ et $0 \geqslant y \geqslant -x^2 \geqslant -1$. Donc K est compacte, et comme f est continue car polynomiale, le problème admet une solution.

2. La contrainte est dessinée en Figure 4.

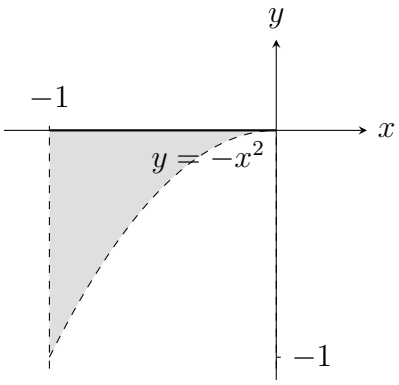


FIGURE 4. Représentation de l'ensemble $K \subset \mathbb{R}^2$.

3. On introduit

$$g_1(x,y) = -x - 1, \qquad g_2(x,y) = y, \qquad g_3(x,y) = -y - x^2, \qquad g_4(x,y) = x,$$

et $G(x) := (g_1(x), g_2(x), g_3(x), g_4(x))^T$. On calcule

$$\nabla g_1(x,y) = -\nabla g_4(x,y) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \quad \nabla g_2(x,y) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad \nabla g_3(x,y) = -\begin{pmatrix} 2x \\ 1 \end{pmatrix}.$$

On vérifie la qualification en chaque point $(x,y) \in K$, $(x,y) \neq (0,0)$. Si une seule contrainte est saturée (et c'est le cas pour g_1, g_2, g_4), le gradient est non nul. Si deux contraintes sont saturées :

- En $(-1,0)$, g_1 et g_2 sont saturées, on choisit $\omega = (1,-1)$ (obtenu en regardant une direction qui permet de rester dans K sur le dessin) et on obtient $\langle \nabla g_i(-1,0), \omega \rangle < 0$ pour $i = 1, 2$.
- En $(-1,-1)$, g_1 et g_3 sont saturées, on prend $\omega = (1,1)$ et on obtient aussi la qualification.

Le seul point non qualifié est $(0,0)$ car 3 contraintes y sont saturées (g_2, g_3, g_4) et il n'existe pas de $\omega \in \mathbb{R}^2$ tel que $\langle \nabla g_i(0,0), \omega \rangle < 0$ pour tout $i \in I((0,0)) = \{2,3,4\}$. En effet, $\nabla g_2(0,0) = \binom{0}{1} = -\nabla g_3(0,0)$ et $\nabla g_4(0,0) = \binom{1}{0}$, donc si $\omega = \binom{\omega_1}{\omega_2}$ est tel que $\langle \nabla g_i(0,0), \omega \rangle < 0$ pour tout $i \in= \{2,3,4\}$, alors $\omega_1 < 0$, $\omega_2 > 0$ et $\omega_2 < 0$, ce qui est impossible.

4. On cherche les points vérifiant les conditions de KKT. Il existe $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \geqslant 0$ tels que :

$$\begin{cases} 1 - \lambda_1 - 2x\lambda_3 + \lambda_4 = 0, \\ -2 + \lambda_2 - \lambda_3 = 0, \\ \lambda_1(-x-1) = \lambda_2 y = \lambda_3(-y-x^2) = \lambda_4 x = 0. \end{cases} \tag{48}$$

On distingue plusieurs cas :

- Si aucune contrainte n'est saturée, alors $g_i(x) \neq 0$ et $g_i(x)\lambda_i = 0$ donne $\lambda_i = 0$ pour tout $i \in \{1,2,3,4\}$. La première équation devient alors $1 = 0$, donc aucun point de l'intérieur de K ne vérifie les conditions KKT.
- Considérons qu'une seule constrainte est qualifiée. Si c'est $i = 1$, alors $\lambda_2 = \lambda_3 = \lambda_4 = 0$ et la seconde équation devient $2 = 0$. Il se passe aussi une contradiction pour $i \in \{2,4\}$. Si $i = 3$ est la seule constrainte saturée, alors le système devient équivalent à $0 = 1 - 2x\lambda_3$ et $\lambda_3 = -2$, mais $\lambda_3 \geqslant 0$ n'est pas respectée. Lorsqu'une seule contrainte est saturée, aucun point ne respecte les conditions de KKT.
- Il ne reste que deux points à vérifier, qui correspondent à deux contraintes saturées.
  - En $(-1,0)$: g_1, g_2 saturées, le système (48) devient $\lambda_1 = 1, \lambda_2 = 2$, donc le point est admissible.
  - En $(-1,-1)$: g_1, g_3 saturées, système impossible.

Le seul point admissible (i.e. qui vérifie les conditions de KKT) est donc $(-1,0)$.

5. Il y a un point non qualifié et parmi les points qualifié, un seul qui est admissible. On compare les valeurs,

$$f(-1,0) = -1, \quad f(0,0) = 0.$$

Le minimum est donc atteint en $(-1,0)$ avec $f(-1,0) = -1$.

Appendix A. Distance relative

On veut comparer deux vecteurs $x, y \in \mathbb{R}^d$, par exemple deux nombres quand $d = 1$. Pour tout $x, y \in \mathbb{R}^d$, définissons leur distance relative

$$\mathbb{D}(x,y) := \frac{\|x-y\|}{\frac{1}{2}\big(\|x\| + \|y\|\big)},$$

à comparer avec la distance "classique" $\|x-y\|$.

Si on a par exemple $x = 0.0001$ et $y = 0.000001$, y est en fait 100 fois plus petit que x mais $|x - y| \simeq 10^{-4}$ qui est petit donc la distance classique ne permet pas de montrer que x et y sont très différents. En revanche $\mathbb{D}(x, y) \simeq 2$ et on voit que la distance relative suggère que x et y sont très différents.

Inversement, si $x = 10010000$ et $y = 10000000$, $|x - y| = 10^4$ qui est grand alors que $\mathbb{D}(x, y) \simeq 10^{-3}$ et on voit que la distance relative suggère que x et y sont en fait proches.

On privilégiera souvent la distance relative quand on regarde des erreurs dans des algorithmes.

Appendix B. Quelques preuves

B.1. **Preuve du Lemme 4.3.**

Preuve du Lemme 4.3. Comme A est symétrique définie positive, il existe une matrice orthogonale Q et une matrice diagonale $D = \mathrm{diag}(d_1, \dots, d_n)$ avec $d_i > 0$ telles que $A = Q^\top DQ$. Posons $y = Qx$. Alors $\|x\| = \|y\|$, $\langle Ax, x\rangle = \langle Dy, y\rangle = \sum_{i=1}^n d_i y_i^2$, et $\langle A^{-1}x, x\rangle = \langle D^{-1}y, y\rangle = \sum_{i=1}^n d_i^{-1} y_i^2$.

Si $x = 0$ l'inégalité est triviale ; supposons donc $x \neq 0$ et normalisons en posant

$$p_i := \frac{y_i^2}{\|y\|^2}, \qquad i = 1, \dots, n.$$

Alors $p_i \geqslant 0$, $\sum_{i=1}^n p_i = 1$, et

$$\frac{\langle Ax, x\rangle}{\|x\|^2} = \sum_{i=1}^n d_i p_i, \qquad \frac{\langle A^{-1}x, x\rangle}{\|x\|^2} = \sum_{i=1}^n d_i^{-1} p_i.$$

Il suffit donc de montrer que pour toute distribution de probabilités (p_i) et pour tout $d_i \in [\lambda_1, \lambda_d]$,

$$\left(\sum_{i=1}^n d_i p_i\right)\left(\sum_{i=1}^n d_i^{-1} p_i\right) \leqslant \frac{(\lambda_d + \lambda_1)^2}{4\lambda_d \lambda_1}.$$

Considérons la fonction sur le simplexe

$$F(p) := \left(\sum_{i=1}^n d_i p_i\right)\left(\sum_{i=1}^n d_i^{-1} p_i\right).$$

c'est-à-dire où $p_i \geqslant 0$, et $\sum_{i=1}^d p_i = 1$. Un argument d'optimisation utilisant le théorème KKT montre que le maximum est atteint lorsque p est porté par au plus deux valeurs $a, b \in [\lambda_1, \lambda_d]$. En effet, introduisons

$$u := \sum_{i=1}^d d_i p_i, \qquad v := \sum_{i=1}^d d_i^{-1} p_i,$$

de sorte que $F(p) = uv$. Pour maximiser F, on impose la contrainte $\sum p_i = 1$ avec multiplicateur de Lagrange μ :

$$L(p, \mu) = uv - \mu\left(\sum_{i=1}^d p_i - 1\right).$$

La condition stationnaire $\frac{\partial L}{\partial p_i} = 0$ donne

$$\frac{\partial L}{\partial p_i} = d_i v + d_i^{-1} u - \mu = 0 \quad \text{si } p_i > 0.$$

Autrement dit, pour tout indice i tel que $p_i > 0$, on doit avoir

$$d_i v + d_i^{-1} u = \mu.$$

Cette équation est linéaire en d_i et d_i^{-1}, qui doit être satisfaite simultanément par tous les d_i associés à des $p_i > 0$. Mais la courbe $g(t) := tv + t^{-1}u$ en fonction de $t > 0$ n'est pas linéaire : c'est une fonction strictement convexe de t. Une équation $g(t) = \mu$ peut avoir au plus deux solutions en t. Ainsi, le support de p (les indices i tels que $p_i > 0$) ne peut contenir plus de deux valeurs distinctes de d_i.

Autrement dit, maximiser F revient à maximiser S avec $0 \leqslant t \leqslant 1$ et $a, b \in [\lambda_1, \lambda_d]$, où

$$S(t; a, b) := (ta + (1-t)b)\,(ta^{-1} + (1-t)b^{-1}) = 1 + t(1-t)\frac{(a-b)^2}{ab}.$$

Ainsi $S(t; a, b)$ est maximal pour $t = \frac{1}{2}$, donnant

$$S(t; a, b) \leqslant S\left(\tfrac{1}{2}; a, b\right) = \frac{(a+b)^2}{4ab} \leqslant \frac{(\lambda_d + \lambda_1)^2}{4\lambda_d \lambda_1}.$$

En revenant aux quantités initiales et en multipliant par $\|x\|^4$, on obtient

$$\langle Ax, x\rangle \left\langle A^{-1}x, x\right\rangle \leqslant \frac{(\lambda_d + \lambda_1)^2}{4\lambda_d \lambda_1} \|x\|^4.$$

□

References

[1] J.-C. Culioli, *Introduction à l'optimisation*, Ellipses, 2012.

Laboratoire AGM - UMR 8088, CNRS - CY Cergy Paris Université, 2 avenue Adolphe Chauvin, Cergy-Pontoise, 95302 cedex, France

Email address: louis.garrigue@cyu.fr