Optimisation numérique

Luca Nenna*

2024-2025

Table des matières

Table des matières **1**

1 Existence, unicité, convexité **5**
1.1 Existence 6
1.2 Notions de calcul différentiel 7
1.3 Convexité et condition nécessaire et suffisante d'optimalité 8
1.4 Stricte convexité et unicité du minimiseur 11

2 Descente de gradient à pas optimal **12**
2.1 Méthode de descente 12
2.2 Descente de gradient à pas optimal 13

3 Descente de gradient à pas optimal II **17**
3.1 Forte convexité et stabilité du minimum 17
3.2 Vitesse de convergence du gradient à pas optimal 19
3.3 Une condition suffisante pour la forte convexité 20

4 Descente de gradient préconditionné à rebroussement **22**
4.1 Choix du pas par rebroussement 22
4.2 Convergence de l'algorithme de descente de gradient préconditionné à rebroussement 24

*Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France. e-mail : `luca.nenna@universite-paris-saclay.fr`

5 Méthode de Newton **26**
5.1 Méthode de Newton pure 26
5.2 Méthode de Newton amortie 29

6 Projection et gradient projeté **32**
6.1 Projection sur un convexe fermé 33
6.2 Condition d'optimalité pour l'optimisation sous contraintes 35

7 Optimisation avec contraintes d'inégalités **39**
7.1 Méthode de pénalisation 40
7.2 Théorème de Karush-Kush-Tucker 42

Bibliographie **49**

Introduction

Motivation Dans de nombreuses applications, la formulation naturelle du problème qu'on cherche à résoudre est un problème d'optimisation :

— Dans la méthode des moindres carrés, on remplace un système linéaire $Ax = b$ surdéterminé et/ou n'ayant pas de solution (par exemple car certaines des égalités se contredisent) par le problème d'optimisation suivant :

$$\min_{x \in \mathbb{R}^N} \|Ax - b\|^2. \tag{1}$$

Le minimiseur x^* de ce problème vérifie "au mieux" la famille d'équation $Ax = b$. Au contraire, lorsqu'un système linéaire $Ax = b$ admet plusieurs solutions, on peut en sélectionner une en considérant le problème

$$\min_{x \in K} \|x\|^2 \quad K = \{x \in \mathbb{R}^N \mid Ax = b\} \tag{2}$$

— Si $\bar{x} \in \mathbb{R}^N$ représente un signal 1D échantilloné avec du bruit ($\bar{x}_i$ représentant par exemple la mesure effectuée en un temps t_i), on peut débruiter le signal en considérant le problème d'optimisation suivant, où $\lambda > 0$ est un paramètre :

$$\min_{x \in \mathbb{R}^N} \|x - \bar{x}\|^2 + \lambda \sum_{1 \leq i \leq N-1} |x_{i+1} - x_i|^2 \tag{3}$$

un compromis entre deux comportements : x^* doit être proche de $\bar{x}$ (c'est le rôle du premier terme $\|x - \bar{x}\|^2$ de la fonction optimisée) mais doit également être "régulier", au sens où deux valeurs successives x_i et x_{i+1} doivent être proches (second terme $\sum_i |x_{i+1} - x_i|^2$).

— En finance, on peut considérer le problème de l'optimisation de portefeuille. Étant donnés N actifs, il s'agit de déterminer le pourcentage $x_i \geq 0$ du portefeuille que l'on investit dans l'actif i. Comme on souhaite investir 100% du portefeuille, ce problème d'optimisation est accompagné d'une contrainte $\sum_{1 \leq i \leq N} x_i = 1$. On pourra donc considérer des problèmes d'optimisation *avec contraintes* de la forme

$$\min_{x \in \Delta} f(x) \text{ où } \Delta = \{x \in \mathbb{R}^N \mid \forall i, x_i \geq 0 \text{ et } \sum_i x_i = 1\}. \tag{4}$$

La fonction f est typiquement de la forme $f(x) = \frac{1}{\varepsilon} |\langle c|x\rangle - r|^2 + \langle x|Qx\rangle$: $c \in \mathbb{R}^N$ représente le rendement des actifs et le premier terme cherche à fixer

le niveau de rendement $\langle c|x\rangle = \sum_i c_i x_i$ à r. Le second terme de la fonction $\langle x|Qx\rangle$ est une mesure de risque : Q est une matrice symétrique mesurant les corrélations entre actifs, et on cherche un investissement minimisant cette corrélation.
— En apprentissage automatique (*machine learning*), de nombreux problèmes peuvent être formulés comme des problèmes d'optimisation. Nous verrons par exemple des problèmes de classification, que l'on résoudra par régression logistique ou par machine à vecteurs support (*support vector machine*).

Pour plus d'exemple, on renvoie au livre de Boyd et Vanderberghe, qui est disponible gratuitement (en anglais) en ligne : `http://web.stanford.edu/~boyd/cvxbook/`.

Problèmes avec/sans contrainte On peut séparer les problèmes en deux grandes classes. Il y a d'une part les problèmes d'optimisation sans contraintes, où l'on cherche à minimiser une fonctionnelle sur $\mathbb{R}^d$ ou sur son domaine de définition *ouvert* (les problèmes (1), (3) et la régression logistique sont de ce type). D'autre part, les problèmes d'optimisation avec contraintes, où l'on cherche à minimiser sur l'ensemble des points de $\mathbb{R}^N$ vérifiant un certain nombre de contraintes d'égalité ou d'inégalité (les problèmes (2), (4) et les machines à vecteurs support sont de ce type).

Convexité Tout les algorithmes et exemples présentés dans ce cours relèvent de l'optimisation *convexe*, où aussi bien la fonction optimisée que le domaine d'optimisation sont supposés convexes. La raison fondamentale pour laquelle on se restreint à ce cas est que pour les problèmes d'optimisation convexe, un minimiseur local est *automatiquement* minimiseur global.

Chapitre 1

Existence, unicité, convexité

Contents

1.1 Existence 6
1.2 Notions de calcul différentiel 7
1.3 Convexité et condition nécessaire et suffisante d'optimalité 8
1.4 Stricte convexité et unicité du minimiseur 11

Dans cette première partie, on s'intéresse à un problème de minimisation d'une fonction $f \in \mathcal{C}^1(\mathbb{R}^d)$:

$$\inf_{x \in \mathbb{R}^d} f(x) \tag{P}$$

Définition 1. On appelle :

(i) *infimum* ou *valeur du problème* de (P) la valeur $\inf_{\mathbb{R}^d} f$.

(ii) *minimiseur global* (ou simplement minimiseur) de (P) tout élément $x^* \in \mathbb{R}^d$ vérifiant $f(x^*) = \inf_{\mathbb{R}^d} f$. On note $\arg\min_{\mathbb{R}^d} f$ l'ensemble des minimiseurs de f (qui peut être vide), i.e.

$$\arg\min_{\mathbb{R}^d} f = \{x \in \mathbb{R}^d \mid f(x) = \inf_{\mathbb{R}^d} f\}.$$

(iii) On appelle *suite minimisante* pour (P) toute suite $x^{(0)}, \dots, x^{(k)}, \dots$ d'éléments de $\mathbb{R}^d$ telle que $\lim_{k \to +\infty} f(x^{(k)}) = \inf_{x \in \mathbb{R}^d} f(x)$.

Remarque 1. Il est possible que le problème (P) n'admette pas de minimiseur : penser par exemple à $f(x) = \exp(x)$ sur $\mathbb{R}$.

Lemme 1. *Il existe une suite minimisante pour le problème* (P).

Démonstration. Par définition de l'infimum, pour tout $k > 0$, il existe un élément $x^{(k)} \in \mathbb{R}^d$ tel que $\inf_{\mathbb{R}^d} f \leq f(x^{(k)}) \leq \inf_{\mathbb{R}^d} f + \frac{1}{k}$, soit $\lim_{k \to +\infty} f(x^{(k)}) = \inf_{\mathbb{R}^d} f$. $\square$

1.1 Existence

Proposition 2. *Soit $f \in \mathcal{C}^0(\mathbb{R}^d)$. On suppose de plus qu'il existe $x_0 \in \mathbb{R}^d$ tel que le sous-niveau $S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x_0)\}$ est compact. Alors le problème d'optimisation (P) admet un minimiseur global.*

Démonstration. Si $f(x_0) = \inf_{\mathbb{R}^d} f$ on a déjà l'existence d'un minimum, à savoir le point x_0 lui-même. On suppose donc maintenant que $f(x_0) > \inf_{\mathbb{R}^d} f$. Soit $(x^{(k)})_{k\geq 0}$ une suite minimisante, qui vérifie donc $\lim_{k\to+\infty} f(x^{(k)}) = \inf_{\mathbb{R}^d} f < f(x_0)$. Alors, pour k suffisamment grand, on a $f(x^{(k)}) \leq f(x_0)$, soit $x^{(k)} \in S$. Comme l'ensemble S est compact, on peut extraire une sous-suite $(x^{(\sigma(k))})_{k\geq 0}$ qui converge vers un point $x^\infty \in S$. Alors, par continuité de f et par définition d'une suite minimisante on a $f(x^\infty) = \lim_{k\to+\infty} f(x^{(\sigma(k))}) = \inf_{\mathbb{R}^d} f$, et x^∞ minimise donc f sur $\mathbb{R}^d$. □

Corollaire 3. *Soit $f \in \mathcal{C}^0(\mathbb{R}^d)$ vérifiant*

$$\lim_{\|x\|\to+\infty} f(x) = +\infty,$$

c'est-à-dire que

$$\forall L \in \mathbb{R}, \exists R \geq 0 \text{ t.q. } \|x\| \geq R \Longrightarrow f(x) \geq L.$$

Alors (P) admet un minimiseur global.

Démonstration. Soit $x_0 \in \mathbb{R}^d$ quelconque et $S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x_0)\}$. Pour montrer l'existence d'un minimiseur, il suffit de démontrer que S est compact. L'ensemble S est fermé comme sous-niveau d'une fonction continue. Supposons S non borné : pour tout k, il existe alors $x^{(k)} \in S$ tel que $\left\|x^{(k)}\right\| \geq qk$. Ainsi, $\lim_{k\to+\infty} x^{(k)} = +\infty$ et $f(x^{(k)}) \leq f(x_0)$, contredisant l'hypothèse. □

Corollaire 4. *Soit $f \in \mathcal{C}^0(\mathbb{R}^d)$ et vérifiant la propriété suivante :*

$$\forall x \in \mathbb{R}^d, f(x) \geq C\|x\|^p + D,$$

où $C > 0, D \in \mathbb{R}$ et $p > 0$. Alors le problème (P) admet un minimiseur.

Démonstration. Soit $x_0 \in \mathbb{R}^d$ quelconque et $S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x_0)\}$. Pour pouvoir appliquer le théorème précédent, il suffit de démontrer que S est compact. Comme on sait déjà que S est fermé comme sous-niveau d'une fonction continue, il suffit de démontrer que cet ensemble est borné. Or, pour tout $x \in S$, on a

$$C\|x\|^p + D \leq f(x) \leq f(x^{(0)})$$

soit $\|x\|^p \leq E := \left|f(x^{(0)}) - D\right|/C$. Ainsi, l'ensemble S est contenu dans la boule centrée en 0 et de rayon $\sqrt[p]{E}$ et est donc borné. $\square$

1.2 Notions de calcul différentiel

Définition 2 (Différentiabilité)**.** Soit $f : \Omega \subset V \to W$ avec Ω ouvert. On dit que f est différentiable en $x_0 \in \Omega$ si et seulement si il existe une application linéaire continue $L \in \mathcal{L}(V, W)$ telle que

$$f(x_0 + h) \underset{h \to 0}{=} f(x_0) + L(h) + o(\|h\|).$$

L'application linéaire L est notée $\mathrm{d}_{x_0} f \in \mathcal{L}(V, W)$ et elle est dite différentielle de f en x_0

L'application f est alors $\mathcal{C}^1(\Omega)$ si elle est différientable dans tout point x dans Ω et l'application

$$\begin{array}{rccl} \mathrm{d}f : & \Omega & \to & ((V, W), \|\cdot\|_{\mathrm{op}}) \\ & x_0 & \mapsto & \mathrm{d}_{x_0} f \end{array}$$

est continue.

Définition 3 (Dérivée directionnelle)**.** Soit $f : \Omega \subset V \to W$ avec Ω ouvert, $x_0 \in \Omega$ et $h \in V$. Quand elle existe, la limite

$$\lim_{t \to 0} \frac{f(x_0 + th) - f(x_0)}{t}$$

est appelée dérivée directionnelle de f dans la direction

Remarque 2. Si f est différentiable en x_0 alors elle admet une dérivée directionnelle dans toute direction $h \in V$ et

$$\mathrm{d}_{x_0} f(h) = \lim_{t \to 0} \frac{f(x_0 + th) - f(x_0)}{t}.$$

La réciproque n'est pas vraie en général ! Par exemple

$$f(x, y) = \begin{cases} \frac{y^2}{x} \text{ if } x \neq 0 \\ y \text{ if } x = 0. \end{cases}$$

Définition 4. Soit $f \in \mathcal{C}^1(\mathbb{R}^d)$. Soit $\langle \cdot, \cdot \rangle$ le produit scalaire usuel il existe un unique vecteur de V, noté $\nabla f(x_0)$ et appelé *gradient* de f en x_0, tel que

$$\mathrm{d}_{x_0} f(h) = \langle \nabla f(x_0), h \rangle \; \forall h \in V.$$

Soit $(e_i)_{1 \leq i \leq d}$ la base canonique de $\mathbb{R}^d$, on a

$$\nabla f(x) = \left(\frac{\partial f}{\partial e_i}(x) \right)_{1 \leq i \leq d} \quad \text{où } \frac{\partial f}{\partial e_i}(x) = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon}(f(x + \varepsilon e_i) - f(x)).$$

Dans la suite on prendra toujours $V = \mathbb{R}^d$ et $W = \mathbb{R}$.

Exemple 1. Considérons $f(x) = \|x\|^2$ sur $\mathbb{R}^d$. En développant le carré de la norme, on obtient $f(x + v) = \|x\|^2 + \langle 2x|v \rangle + \|v\|^2$. On en déduit que $\nabla f(x) = 2x$, ce qui est conforme avec le calcul.

On dit que $f : \Omega \subset V \to W$ est deux fois différentiable en x_0 si elle est différentiable dans un voisinage de x_0 et $\mathrm{d}f$ est différentiable en x_0. On note cette dérivée $\mathrm{d}^2_{x_0}$ qui est un élément de $\mathcal{L}(V, \mathcal{L}(V, W))$. En particulier on a que $\mathrm{d}^2_{x_0} f(h, k)$ est la dérivée directionnelle de $x \mapsto \mathrm{d}_x f(h)$ dans la direction k.

Définition 5. Soit $f \in \mathcal{C}^2(\Omega)$, où $\Omega \subseteq \mathbb{R}^d$ est un ouvert. On appelle *hessienne* de f en $x_0 \in \Omega$ la matrice associée a la forme bilinéaire $\mathrm{d}^2_{x_0}$ dans la base canonique. En particulier

$$\mathrm{D}^2 f(x) = \left(\frac{\partial^2 f}{\partial e_i \partial e_j}(x) \right)_{1 \leq i,j \leq d},$$

où l'on a noté $(e_i)_{1 \leq i \leq d}$ la base canonique de $\mathbb{R}^d$. et où

1.3 Convexité et condition nécessaire et suffisante d'optimalité

1.3.1 Convexité

Définition 6. Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ est *convexe* si

$$\forall (x, y) \in \mathbb{R}^d, \forall t \in [0, 1], \; f((1 - t)x + ty) \leq (1 - t)f(x) + tf(y).$$

Exercice 1. Soit $f : \mathbb{R}^d \to \mathbb{R}$ une fonction convexe. Montrer que l'ensemble

$$\{x \in \mathbb{R}^d \mid f(x) \leq C\}$$

est convexe quel que soit C. En déduire que l'ensemble des minimiseurs de (P) est un ensemble convexe fermé (possiblement vide).

Proposition 5. *Soit $f \in \mathcal{C}^1(\mathbb{R}^d)$. Alors les propositions suivantes sont équivalentes :*
(i) f est convexe sur $\mathbb{R}^d$,
(ii) $\forall x, y \in \mathbb{R}^d$, la fonction $g : t \in [0,1] \mapsto f((1-t)x + ty)$ est convexe.
(iii) $\forall x, y \in \mathbb{R}^d, f(y) \geq f(x) + \langle y - x | \nabla f(x) \rangle$,
(iv) $\forall x, y \in \mathbb{R}^d, \langle \nabla f(x) - \nabla f(y) | x - y \rangle \geq 0$.

Lemme 6. *Soit $f \in \mathcal{C}^1(\mathbb{R}^d)$. Étant donnés $x \in \Omega$ et $v \in \mathbb{R}^d$, on définit $x_t = x + tv$ et $g(t) = f(x_t)$. Alors,*

$$g'(t) = \langle \nabla f(x_t) | v \rangle. \tag{1.1}$$

Démonstration de la proposition 5. *(i)$\Longleftrightarrow$(ii)* conséquence directe de la définition.
(ii)$\Longrightarrow$(iii). Soit x, y dans $\mathbb{R}^d$ et $g : t \mapsto f(x_t)$ où $x_t = (1-t)x + tv = x + t(y-x)$, qui est convexe par hypothèse. Par convexité, on a $g(t) \geq g(0) + tg'(0)$, soit par le lemme $f(x) + \langle \nabla f(x) | y - x \rangle + \leq f(y)$.
(iii)$\Longrightarrow$(iv) Il suffit de sommer l'inégalité (iii) et la même inégalité où l'on a inversé le rôle de x et y.
(iv)$\Longrightarrow$(ii) Soit encore $g : t \mapsto f(x_t)$ où $x_t = (1-t)x + tv = x + t(y-x)$. Comme $g'(t) = \langle \nabla f(x_t) | y - x \rangle$, (lemme 6) l'inégalité (iv) appliquée en x_s et x_t (où $t > s$) nous donne

$$g'(t) - g'(s) = \langle \nabla f(x_t) - \nabla f(x_s) | y - x \rangle = \frac{1}{t-s} \langle \nabla f(x_t) - \nabla f(x_s) | x_t - x_s \rangle \geq 0,$$

et g' est donc croissante sur $[0,1]$. Ainsi, g est convexe. □

Lemme 7 (Taylor-Lagrange)**.** *Soit $f \in \mathcal{C}^2(\mathbb{R}^d)$, $x, v \in \mathbb{R}^d$ et $x_t = x + tv$. Alors,*

$$\forall t \geq 0, \exists s \in [0,t], \quad f(x_t) = f(x) + t\langle \nabla f(x) | v \rangle + \frac{t^2}{2} \langle D^2 f(x_s) v | v \rangle.$$

Lemme 8. *Soit $\Omega \subseteq \mathbb{R}^d$ ouvert, $f \in \mathcal{C}^2(\Omega)$, $x \in \Omega$ et $v \in \mathbb{R}^d$ et $g : t \mapsto f(x_t)$ où $x_t = x + tv$. Alors,*

$$g''(t) = \langle \mathrm{D}^2 f(x_t) v | v \rangle. \tag{1.2}$$

Démonstration. On fait le calcul en coordonnées, en notant $(e_k)_k$ la base canonique :

$$g(t) = f(x + \sum_k t v_k e_k)$$

$$g'(t) = \sum_i v_i \frac{\partial f}{\partial e_i} \left(x + \sum_k t v_k e_k \right) = \langle \nabla f(x_t) | v \rangle$$

$$g''(t) = \sum_i \sum_j v_i v_j \frac{\partial^2 f}{\partial e_j \partial e_i}\left(x + \sum_k t v_k e_k\right) = \langle \mathrm{D}^2 f(x_t) v | v \rangle \qquad \square$$

Proposition 9. *Soit $f \in \mathcal{C}^2(\mathbb{R}^d)$. Si $\forall x \in \mathbb{R}^d, \mathrm{D}^2 f(x) \succeq 0$, alors f est convexe.*

Démonstration de la proposition 9. Considérons $x, y \in \Omega$ et $g(t) = f(x_t)$ où $x_t = (1-t)x + ty$. Alors, $g''(t) = \langle \mathrm{D}^2 f(x_t)(y-x)|y-x\rangle$ est positif par hypothèse, de sorte que par Taylor-Lagrange

$$g(1) = g(0) + g'(0) + \frac{s^2}{2} g''(s) \geq g(0) + g'(0),$$

soit $f(y) \geq f(x) + \langle \nabla f(x)|y-x\rangle$. La proposition 5 montre que f est convexe. $\square$

1.3.2 Condition nécessaire d'optimalité

Théorème 10 (Fermat)**.** *Soit $f \in \mathcal{C}^1(\mathbb{R}^d)$, et x^* un minimiseur de (P). Alors*

$$\nabla f(x^*) = 0. \quad \text{(\textit{équation d'Euler })}$$

Remarque 3. La contraposée est fausse : prendre $f(x) = x^3$ sur $\Omega = \mathbb{R}$: le point 0 vérifie $f'(0) = 0$ mais n'est pas un minimiseur (même local).

Démonstration. Si x^* est un minimiseur de f sur Ω, on a pout tout $\varepsilon \in \mathbb{R}$, $f(x^* + \varepsilon e_i) \geq f(x^*)$. Ainsi,

$$\forall 0 < \varepsilon \leq \varepsilon_0, \quad \frac{f(x^* + \varepsilon e_i) - f(x^*)}{\varepsilon} \geq 0.$$

En passant à la limite, on obtient $\frac{\partial f}{\partial e_i}(x^*) \geq 0$. De même, en considérant le cas $\varepsilon < 0$

$$\forall -\varepsilon_0 \leq \varepsilon < 0, \quad \frac{f(x^* + \varepsilon e_i) - f(x^*)}{\varepsilon} \leq 0,$$

d'où l'on tire en passant à la limite $\frac{\partial f}{\partial e_i}(x^*) \leq 0$, soit *in fine* $\frac{\partial f}{\partial e_i}(x^*) = 0$. $\square$

1.3.3 Condition suffisante d'optimalité

Théorème 11. *Soit $\Omega \subseteq \mathbb{R}^d$ un ouvert convexe et $f \in \mathcal{C}^1(\Omega)$ convexe. Alors $x^* \in \Omega$ est un minimiseur de (P) si et seulement si $\nabla f(x^*) = 0$.*

Démonstration. Le théorème de Fermat nous donne déjà le sens direct. Pour la réciproque, il suffit de remarquer que si $\nabla f(x^*) = 0$, la proposition précédente donne

$$\forall y \in \Omega, f(y) \geq f(x^*) + \langle y - x^* | \nabla f(x^*) \rangle = f(x^*),$$

de sorte que x^* est bien un minimiseur de (P). □

1.4 Stricte convexité et unicité du minimiseur

Définition 7. Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ est strictement convexe si

$$\forall x \neq y \in \mathbb{R}^d, \forall t \in]0,1[, \;\; f((1-t)x + ty) < (1-t)f(x) + tf(y).$$

Proposition 12. *Soit $f : \mathbb{R}^d \to \mathbb{R}$ strictement convexe. Alors f admet au plus un minimiseur sur $\mathbb{R}^d$.*

Démonstration. Par l'absurde, supposons que f admette deux minimiseurs distincts $x^* \neq y^* \in \mathbb{R}^d$. Alors, par stricte convexité de la fonction f on a $f(z^*) < \frac{1}{2}(f(x^*) + f(y^*)) = f(x^*)$, où $z^* = \frac{1}{2}(x^* + y^*)$, contredisant l'hypothèse que x^* minimise f. □

Remarque 4. Cette proposition ne dit rien de l'existence d'un minimiseur.

Chapitre 2

Descente de gradient à pas optimal

On souhaite résoudre numériquement le problème de minimisation d'une fonction $f \in \mathcal{C}^0(\mathbb{R}^d)$. Comme en général il n'est pas raisonnable d'espérer calculer de manière exacte un minimiseur ou même la valeur de l'infimum du problème (P), on cherchera à l'*approcher*. Il s'agira de construire une suite $(x^{(k)})_{k\geq 0}$ de points vérifiant une des deux propriétés suivantes :

(a) la suite $x^{(k)}$ est minimisante pour (P), i.e. $\lim_{k\to+\infty} f(x^{(k)}) = \inf_{\mathbb{R}^d} f$.

(b) la suite $x^{(k)}$ converge vers un minimiseur de f sur $\mathbb{R}^d$.

Dans ce chapitre, nous nous intéresserons à la construction de suites $x^{(k)}$ vérifiant la seconde propriété (qui est bien sûr plus forte que la première).

2.1 Méthode de descente

Vocabulaire On appelle *méthode de descente* un procédé algorithmique permettant de construire itérativement une suite vérifiant (a) ou (b). Typiquement, une méthode de descente prend la forme suivante

$$\begin{cases} d^{(k)} = \dots & \textit{direction de descente} \\ t^{(k)} = \dots & \textit{pas de descente} \\ x^{(k+1)} = x^{(k)} + t^{(k)} d^{(k)} \end{cases}$$

Un tel algorithme est appelé méthode de descente si $f(x^{(k+1)}) \leq f(x^{(k)})$. Dans ce cours, on considèrera les possibilités suivantes :

(a) La *direction de descente* peut être égale à l'opposé du gradient, $d^{(k)} = -\nabla f(x^{(k)})$, auquel cas on parle de méthode de descente de gradient. Lorsque f est $\mathcal{C}^2$, il peut être plus avantageux de choisir $d^{(k)} = -\mathrm{D}^2 f(x^{(k)})^{-1} \nabla f(x^{(k)})$, auquel cas on parle de méthode de Newton.

(b) Le *pas de descente* peut être choisi constant ($t^{(k)} = \tau$), optimal (cf (2.1)), ou obtenu par des constructions un peu plus complexes, permettant de garantir la convergence de la méthode.

Définition 8. Soit $f \in \mathcal{C}^0(\mathbb{R}^d)$. On appelle *direction de descente* en $x \in \mathbb{R}^d$ tout vecteur v tel que $\exists \tau > 0, \forall t \in [0, \tau], f(x + tv) < f(x)$.

Exercice 2. Si $f \in \mathcal{C}^1(\mathbb{R}^d)$ et $\langle v | \nabla f(x) \rangle < 0$, alors v est une direction de descente.

Si f est différentiable en x, on a $f(x+tv) = f(x) + t\langle \nabla f(x) | v \rangle + \mathrm{o}(t)$. On cherche naturellement une direction de descente rendant le produit scalaire $\langle \nabla f(x) | v \rangle$ le plus petit possible, menant au choix de $d^{(k)} = -\nabla f(x^{(k)})$:

Lemme 13. *Soit* $f \in \mathcal{C}^1(\mathbb{R}^d)$ *et* $x_0 \in \mathbb{R}^d$. *Alors,* $\min_{\|v\|=1} \mathrm{d}_{x_0} f(v) = -\|\nabla f(x_0)\|$ *et si* $\nabla f(x_0) \neq 0$, *l'unique minimiseur est* $v = -\nabla f(x_0) / \|\nabla f(x_0)\|$.

Démonstration. On a $\mathrm{d}_{x_0} f(v) = \langle \nabla f(x_0) | v \rangle \geq -\|\nabla f(x_0)\| \|v\|$ par Cauchy-Schwarz, avec égalité si et seulement si v est positivement homogène à $-\nabla f(x_0)$. Comme $\|v\| = 1$, on a $v = -\nabla f(x_0) / \|\nabla f(x_0)\|$. □

2.2 Descente de gradient à pas optimal

Définition 9. Soit $f \in \mathcal{C}^1(\mathbb{R}^d)$. L'algorithme de descente de gradient à pas optimal est donné par :

$$\begin{cases} d^{(k)} = -\nabla f(x^{(k)}) \\ t^{(k)} \in \arg\min_t f(x^{(k)} + td^{(k)}) \\ x^{(k+1)} = x^{(k)} + t^{(k)} d^{(k)} \end{cases} \tag{2.1}$$

Remarque 5. Par construction, les itérées vérifient

$$f(x^{(k+1)}) \leq f(x^{(k)})$$
$$\langle \nabla f(x^{(k+1)}) | \nabla f(x^{(k)}) \rangle = 0.$$

Pour la deuxième égalité, il suffit de remarquer que si l'on pose $g(t) = f(x^{(k)} + td^{(k)})$, alors par définition de $t^{(k)}$,

$$g'(t^{(k)}) = 0 = \langle \nabla f(x^{(k)} + t^{(k)} d^{(k)}) | d^{(k)} \rangle = \langle \nabla f(x^{(k+1)}) | \nabla f(x^{(k)}) \rangle.$$

Remarque 6. Pour pouvoir mettre en œuvre cet algorithme il faut pouvoir calculer le pas optimal $t^{(k)}$ à chaque itération, ce qui implique de résoudre un problème d'optimisation (sur $\mathbb{R}$). Ceci n'est faisable de manière exacte que dans un nombre très limité de cas. En général, on préfèrera d'autre méthode de calcul du pas.

Exemple 2. Soit $f : x \in \mathbb{R}^d \mapsto \frac{1}{2}\langle Qx|x\rangle + \langle b|x\rangle$ où Q est une matrice symétrique définie positive. Alors f est strictement convexe et $\nabla f(x) = Qx + b$. Soit $x^{(k)} \in \mathbb{R}^d$, $d^{(k)} = -\nabla f(x^{(k)})$. Pour calculer le pas $t^{(k)} \in \mathbb{R}$, on cherche le minimum de $g : t \in \mathbb{R} \mapsto f(x_t)$ où $x_t = x^{(k)} + td^{(k)}$. La fonction g est convexe et atteint donc son minimum en l'unique point $t^{(k)}$ vérifiant $g'(t^{(k)}) = 0$. Or, $g'(t) = \langle \nabla f(x_t)|d^{(k)}\rangle$, soit

$$\begin{aligned} g'(t^{(k)}) = 0 &\iff \langle Q(x^{(k)} + t^{(k)}d^{(k)}) + b|d^{(k)}\rangle = 0 \\ &\iff t^{(k)}\langle Qd^{(k)}|d^{(k)}\rangle - \langle d^{(k)}|d^{(k)}\rangle = 0. \end{aligned}$$

Ainsi, $t^{(k)} = \langle d^{(k)}|d^{(k)}\rangle / \langle Qd^{(k)}|d^{(k)}\rangle$. En résumé, dans le cas d'une fonction f de la forme considérée, l'algorithme de descente de gradient à pas optimal s'écrit

$$\begin{cases} d^{(k)} = -(Qx^{(k)} + b) \\ t^{(k)} = \frac{\langle d^{(k)}|d^{(k)}\rangle}{\langle Qd^{(k)}|d^{(k)}\rangle} \\ x^{(k+1)} = x^{(k)} + t^{(k)}d^{(k)}. \end{cases} \tag{2.2}$$

On suppose dans la suite que l'ensemble

$$S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x^{(0)})\} \text{ est compact,} \tag{2.3}$$

ce qui garantit (proposition 2) l'existence d'un minimiseur de f sur Ω.

Lemme 14. *Sous l'hypothèse (2.3), le minimum dans la définition du pas est atteint.*

Démonstration. Il s'agit de démontrer que la fonction $g : t \mapsto f(x^{(k)} + td^{(k)})$ atteint son minimum. On suppose que $d^{(k)}$ est non nul : sinon g est constante et atteint évidemment son minimum. Grâce à la proposition 2, il suffit de montrer que le sous-niveau $S_g := \{t \in \mathbb{R} \mid g(t) \leq f(x^{(k)})\}$ est compact. Ce sous-niveau est fermé comme image inverse du fermé $]-\infty, f(x^{(k)})]$ par la fonction continue g. Montrons maintenant que S_g est borné. Pour cela, nous utilisons l'hypothèse que S est compact donc borné : $\exists R \geq 0, \forall x \in S, \|x\| \leq R$. Si $t \in S_g$, alors $x^{(k)} + td^{(k)} \in S$, de sorte que

$$\left\| x^{(k)} + td^{(k)} \right\| \leq R$$

soit

$$|t| \leq \frac{R + \|x^{(k)}\|}{d^{(k)}}.$$

Ainsi S_g est compact et par la proposition 2, g atteint son minimum sur $\mathbb{R}$. $\square$

Théorème 15. *Soit* $f \in \mathcal{C}^2(\mathbb{R}^d)$ *vérifiant*

(i) le sous-niveau $S = \{x \in \mathbb{R}^d \mid f(x) \le f(x^{(0)})\}$ *est compact.*

(ii) $\exists M > 0,\ \forall x \in S,\ \mathrm{D}^2 f(x) \preceq M\mathrm{Id}$,

(iii) f *est strictement convexe,*

Alors les itérées de l'algorithme (2.1) convergent vers l'unique minimiseur global de f *sur* Ω.

On utilisera la proposition suivante :

Proposition 16. *Soit* $(x_k)_{k\ge1}$ *une suite bornée de* $\mathbb{R}^d$ *admettant une unique valeur d'adhérence* $\bar{x}$.[a] *Alors,* $\lim_{k\to+\infty} x_k = \bar{x}$.

a. On rappelle que $\bar{x}$ est valeur d'adhérence de la suite $(x_k)_{k\ge1}$ si et seulement si il existe une suite extraite $(x_{\sigma(k)})_{k\ge1}$ dont la limite est $\bar{x}$

Démonstration du théorème 15. Soit $k \ge 1$ et $g(t) = f(x^{(k)} + td^{(k)}) = f(x^{(k)} - t\nabla f(x^{(k)}))$. Par définition du pas optimal $t^{(k)}$ on a

$$f(x^{(k+1)}) = \min_t f(x^{(k)} + td^{(k)}) = \min_t g(t),$$

et nous allons utiliser un développement de Taylor pour estimer le minimum de g. Par le lemme 8, et en posant $x_t = x^{(k)} + td^{(k)}$ on a

$$g'(t) = \langle \nabla f(x_t) | d^{(k)} \rangle, \qquad g''(t) = \langle \mathrm{D}^2 f(x_t) d^{(k)} | d^{(k)} \rangle.$$

Soit $\sigma = \{t \in \mathbb{R} \mid x_t \in S\}$ et $t \in \sigma$. Par Taylor-Lagrange, pour tout $t \in \sigma$, il existe $s \in [0,t]$ tel que $g(t) = g(0) + tg'(0) + \frac{t^2}{2} g''(s)$. Ainsi,

$$\begin{aligned} g(t) &= g(0) + tg'(0) + \frac{t^2}{2} g''(s) \\ &= f(x^{(k)}) - t \left\| \nabla f(x^{(k)}) \right\|^2 + \frac{t^2}{2} \langle \mathrm{D}^2 f(x_s) \nabla f(x^{(k)}) | \nabla f(x^{(k)}) \rangle \end{aligned}$$

Comme $s \in [0,t]$, alors $s = \gamma t$ avec $\gamma \in [0,1]$ de sorte que $f(x_s) = f((1-\gamma)x_0 + \gamma x_t) \le f(x^{(0)})$. Ainsi, $x_s \in S$. Par hypothèse, on a donc $\mathrm{D}^2 f(x_s) \le M\mathrm{id}$, ce qui donne en utilisant (ii)

$$g(t) \le f(x^{(k)}) + \left(\frac{M}{2} t^2 - t \right) \left\| \nabla f(x^{(k)}) \right\|^2$$

Le minimum de ce second membre est atteint en $t = 1/M$ et on a donc

$$f(x^{(k+1)}) \le \min_t g(t) \le f(x^{(k)}) - \frac{1}{2M} \left\| \nabla f(x^{(k)}) \right\|^2,$$

de sorte que

$$\left\| \nabla f(x^{(k)}) \right\|^2 \le 2M(f(x^{(k)}) - f(x^{(k+1)})). \tag{2.4}$$

Ainsi, pour tout $K \geq 0$ on a

$$\sum_{0 \leq k \leq K} \left\| \nabla f(x^{(k)}) \right\|^2 \leq 2M(f(x^{(0)}) - f(x^{(K)})) \leq 2M(f(x^{(0)}) - \inf_{\mathbb{R}^d} f).$$

La série de terme général $\left\| \nabla f(x^{(k)}) \right\|^2$ est donc convergente, d'où l'on déduit que $\lim_{k \to +\infty} \left\| \nabla f(x^{(k)}) \right\| = 0$.

Montrons enfin que $\lim_{k \to +\infty} x^{(k)} = x^*$, où x^* est l'unique minimum de f sur $\mathbb{R}^d$, l'unicité provenant de la stricte convexité de f et de la proposition 12. Comme $f(x^{(k)}) \leq f(x^{(0)})$, le point $x^{(k)}$ appartient à S, qui est par hypothèse compact donc borné. Pour montrer que la suite $x^{(k)}$ converge vers x^*, il suffit par la proposition 16 de démontrer qu'elle admet x^* pour seule valeur d'adhérence. Soit donc $(x^{(\sigma(k))})$ une sous-suite convergeant vers une valeur d'adhérence $\bar{x} \in S$. Alors, comme $f \in \mathcal{C}^1(\mathbb{R}^d)$, $\nabla f(\bar{x}) = \lim_{k \to +\infty} \nabla f(x^{(\sigma(k))}) = 0$. Par convexité (théorème 11), on sait que $\bar{x}$ est un minimiseur de f sur $\mathbb{R}^d$. Par unicité du minimiseur, on en déduit que $\bar{x} = x^*$. $\square$

Chapitre 3

Descente de gradient à pas optimal II

3.1 Forte convexité et stabilité du minimum

Motivation. Soit $f \in \mathcal{C}^0(\mathbb{R}^d)$ atteignant son minimum x^* sur $\mathbb{R}^d$. On a vu dans la proposition 12 que si f est strictement convexe, alors f admet au plus un minimiseur x^*. Une manière de reformuler cette propriété est la suivante :

$$f(x) \leq f(x^*) \Longrightarrow x = x^*.$$

En pratique, nos algorithmes sont incapables de calculer x^* mais permettent au mieux de calculer une suite $x^{(k)}$ tel que $f(x^{(k)}) \leq f(x^*)+\varepsilon_k$ où $\lim_{k\to+\infty} \varepsilon_k = 0$. On aimerait pouvoir en déduire que $x^{(k)}$ est "proche" de la solution x^*, i.e. $\left\|x^{(k)} - x^*\right\| \leq C\varepsilon_k^\alpha$ pour un certain exposant $\alpha > 0$ et une constante $C > 0$. Nous verrons qu'une telle inégalité est vraie pour $\alpha = \frac{1}{2}$ si la fonction f est *fortement convexe*.

Définition 10. Un sous-ensemble X de $\mathbb{R}^d$ est dit *convexe* s'il contient tout segment reliant deux de ses points, c'est-à-dire :

$$\forall (x, y) \in X, \forall t \in [0, 1],\ (1 - t)x + ty \in X.$$

Définition 11. Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ est dite *m-fortement convexe* sur un ensemble convexe $X \subseteq \mathbb{R}^d$, où $m > 0$, si la fonction $f - \frac{m}{2} \|\cdot\|^2$ est convexe.

Exercice 3. Si f est m-fortement convexe, alors elle est aussi strictement convexe.

Proposition 17. *Soit $f \in \mathcal{C}^1(\mathbb{R}^d)$ et $X \subseteq \mathbb{R}^d$ convexe. Alors les implications suivantes sont vraies (i)$\Longrightarrow$(ii)$\Longleftrightarrow$(iii)$\Longleftrightarrow$(iv), où*

(i) $f \in \mathcal{C}^2(\mathbb{R}^d)$ *et* $\forall x \in X,\ \mathrm{D}^2 f(x) \succeq m\mathrm{Id}$ *;*

(ii) f *est* m*-fortement convexe sur* X *;*

(iii) $\forall x, y \in X, \quad f(y) \geq f(x) + \langle \nabla f(x) | y - x\rangle + \frac{m}{2} \|x - y\|^2$

(iv) $\forall x, y \in X, \quad \langle \nabla f(y) - \nabla f(x) | y - x\rangle \geq m \|x - y\|^2$

Démonstration. La fonction f est m-fortement convexe si et seulement si la fonction $g = f - \frac{m}{2} \|\cdot\|^2$ est convexe. L'équivalence est s'obtient et utilisant les proposition 5 9 pour caractériser la convexité de g. □

Exemple 3. La fonction $f : t \mapsto \exp(t)$ n'est pas fortement convexe sur $\mathbb{R}$, mais elle est m-fortement convexe sur tout segment $[a, b]$ pour $m = \min_{t\in[a,b]} \exp(t)$. De même, fonction $t \mapsto t^4$ n'est pas fortement convexe sur $\mathbb{R}$ mais est fortement convexe sur tout segment ne contenant pas l'origine.

Corollaire 18. *Soit $f \in \mathcal{C}^1(\mathbb{R}^d)$ admettant un minimiseur $x^* \in \mathbb{R}^d$. On suppose que f est m-fortement convexe sur un ensemble convexe S contenant x^*. Alors, pour tout $x \in S$,*

(i) $\|x - x^*\|^2 \leq \frac{2}{m}(f(x) - f(x^*))$

(ii) $\|x - x^*\| \leq \frac{1}{m} \|\nabla f(x)\|$

(iii) $f(x) - f(x^*) \leq \frac{1}{2m} \|\nabla f(x)\|^2$

Remarque 7. Le point (i) montre immédiatement qu'il n'existe pas d'autre minimiseur de f dans S : si x est un autre minimiseur, on a $f(x) = f(x^*)$, de sorte que par (i), $x = x^*$. Mais il faut surtout retenir que si x est “presqu'un minimiseur”, au sens où $f(x) \leq f(x^*) + \varepsilon$, alors x est “proche” du minimiseur x^*, plus précisément $\|x - x^*\| \leq \sqrt{2\varepsilon/m}$. On tire des conclusions similaires du point (ii) : si la condition d'optimalité ($\nabla f(x^*) = 0$) est “presque vérifiée”, i.e. si $\|\nabla f(x)\| \leq \varepsilon$ est petit, alors $\|x - x^*\| \leq \varepsilon/m$.

Exemple 4. Pour $f(t) = t^4$, on a $x^* = 0$ et l'inégalité (i), qui s'écrit $t^2 \leq \frac{2}{m} t^4$, n'est vraie pour aucun $m > 0$.

Démonstration. (i) est une conséquence immédiate de la formulation équivalente de la forte convexité donnée dans la proposition 17.(ii) :

$$f(x) \geq f(x^*) + \langle x - x^* | \nabla f(x^*)\rangle + \frac{m}{2} \|x - x^*\| = f(x^*) + \frac{m}{2} \|x - x^*\|^2,$$

où l'on a utilisé $\nabla f(x^*) = 0$ par optimalité de x^*. (ii) s'obtient de la même manière, en utilisant proposition 17.(iii).

(iii) Par l'inégalité (ii) de la proposition 17, on a pour tout $x, y \in S$

$$f(y) \geq Q(x, y) := f(x) + \langle y - x | \nabla f(x)\rangle + \frac{m}{2} \|y - x\|^2.$$

Ainsi, $f(y) \geq \inf_z g(z)$ où $g(z) = Q(x,z)$. Comme g est convexe sur $\mathbb{R}^d$, son unique minimiseur z^* est solution de l'équation

$$\nabla g(z^*) = 0 = \nabla f(x) + m(z^* - x),$$

c'est-à-dire que $z^* = x - \frac{1}{m}\nabla f(x)$. Ainsi, nous venons de démontrer que

$$\forall y \in S, f(y) \geq g(z^*) = f(x) - \frac{1}{m}\|\nabla f(x)\|^2 + \frac{m}{2}\left\|\frac{1}{m}\nabla f(x)\right\|^2 = f(x) - \frac{1}{2m}\|\nabla f(x)\|^2.$$

En prenant $y = x^*$ dans l'inégalité précédente, on obtient le résultat voulu. □

3.2 Vitesse de convergence du gradient à pas optimal

Définition 12 (Convergence linéaire)**.** Soit $(u_k)_{k\geq 0}$ une suite de limite u^*. La suite (u_k) converge *linéairement* vers u^* s'il existe $0 \leq \kappa < 1$ et $k_0 \in \mathbb{N}$ telle que

$$\forall k \geq k_0, \|u_{k+1} - u^*\| \leq \kappa \|u_k - u^*\|.$$

Théorème 19. *Soit $f \in \mathcal{C}^2(\mathbb{R}^d)$ vérifiant*

(i) le sous-niveau $S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x^{(0)})\}$ est compact.

(ii) $\exists M,\ \forall x \in S,\ \mathrm{D}^2 f(x) \preceq M\mathrm{Id}$,

(iii) $\exists m > 0,\ \forall x \in S,\ \mathrm{D}^2 f(x) \succeq m\mathrm{Id}$,

Alors les itérées de l'algorithme (2.1) convergent vers l'unique minimiseur global de f sur $\mathbb{R}^d$, et de plus, en posant $c = M/m$,

$$f(x^{(k+1)}) - f(x^*) \leq \left(1 - \frac{1}{c}\right)\left(f(x^{(k)}) - f(x^*)\right). \tag{3.1}$$

Remarque 8. En d'autres termes, $(f(x^{(k)}))_{k\geq 0}$ converge linéairement vers $f(x^*)$.

Remarque 9. Si une fonction $f \in \mathcal{C}^2(\mathbb{R}^d)$, vérifie $m\mathrm{Id} \preceq \mathrm{D}^2 f(x) \leq M\mathrm{Id}$, on pourra dira que le *conditionnement* de f est majoré par $c = \frac{M}{m}$. L'inégalité (3.1) montre que cette quantité est cruciale pour comprendre la vitesse de convergence de l'algorithme de descente de gradient. Par exemple, si l'on souhaite estimer $f(x^*)$ à $\varepsilon > 0$ près, d'après l'inégalité (3.1), il suffit d'interrompre l'algorithme de descente de gradient à pas optimal après k itérations où

$$\left(1 - \frac{1}{c}\right)^k (f(x^{(0)}) - f(x^{(*)})) \leq \varepsilon,$$

soit

$$k \geq \frac{\log\left(\frac{f(x^{(0)}) - f(x^{(*)})}{\varepsilon}\right)}{\log\left(\frac{1}{1-1/c}\right)} \sim_{c \to +\infty} c \log\left(\frac{f(x^{(0)}) - f(x^{(*)})}{\varepsilon}\right)$$

Ainsi, dans le cas $c \gg 1$, le nombre d'itération est *proportionnel* au conditionnement !

Démonstration du théorème 19. Le corollaire (18) nous donne

$$f(x^k) - f(x^*) \leq \frac{1}{2m} \left\|\nabla f(x^{(k)})\right\|^2.$$

En combinant avec l'inégalité (2.4), on obtient

$$2m(f(x^k) - f(x^*)) \leq \left\|\nabla f(x^{(k)})\right\|^2 \leq 2M(f(x^{(k)}) - f(x^{(k+1)})),$$

de sorte qu'en posant $c = M/m$, on obtient bien l'inégalité voulue. □

3.3 Une condition suffisante pour la forte convexité

Proposition 20. *Soit $f \in \mathcal{C}^2(\mathbb{R}^d)$ et $S \subseteq \mathbb{R}^d$ compact tel que $\forall x \in S$, $\mathrm{D}^2 f(x)$ est définie positive. Alors $\exists M \geq m > 0$ tels que*

$$\forall x \in S, m\mathrm{Id} \preceq \mathrm{D}^2 f(x) \preceq M\mathrm{Id}.$$

Démonstration. L'ensemble $K = \{(x, v) \in S \times \mathbb{R}^d \mid \|v\| = 1\}$ est compact comme fermé borné. La fonction $(x, v) \in K \mapsto \langle \mathrm{D}^2 f(x) v | v\rangle$ est continue. Elle es donc bornée sur K et atteint ses bornes : ainsi

$$\forall (x, v) \in K, m \leq \langle \mathrm{D}^2 f(x) v | v\rangle \leq M,$$

et il existe $(x_m, v_m) \in K$ tels que $m = \langle \mathrm{D}^2 f(x_m) v_m | v_m\rangle > 0$ par hypothèses et comme $v_0 \neq 0$ (car $\|v_0\| = 1$). On en déduit l'inégalité voulue en remarquant que

$$\forall w \in \mathbb{R}^d, \langle \mathrm{D}^2 f(x) w | w\rangle = \langle \mathrm{D}^2 f(x) \frac{w}{\|w\|} | \frac{w}{\|w\|}\rangle \|w\|^2,$$

de sorte que $m \|w\|^2 \leq \langle \mathrm{D}^2 f(x) w | w\rangle \leq M \|w\|^2$ comme souhaité. □

Cette proposition permet d'énoncer une variante non-quantitative du théorème 19

Corollaire 21. *Soit* $f \in \mathcal{C}^2(\mathbb{R}^d)$ *vérifiant*

(i) le sous-niveau $S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x^{(0)})\}$ *est compact,*

(ii) $\forall x \in S,\ \mathrm{D}^2 f(x)$ *est définie positive.*

Alors les itérées de l'algorithme (2.1) *convergent vers l'unique minimiseur global* x^* *de* f *sur* $\mathbb{R}^d$*, et de plus* $(f(x^{(k)}))_{k \geq 0}$ *converge linéairement vers* $f(x^*)$*.*

Chapitre 4

Descente de gradient préconditionné à rebroussement

Une des objections à l'algorithme de descente de gradient à pas optimal est le calcul du pas demande de résoudre un problème d'optimisation 1D à chaque itération. Nous verrons dans ce chapitre une manière simple de choisir le pas qui donne les même garanties de convergence. Nous allons également un peu relaxer l'hypothèse que $d^{(k)} = -\nabla f(x^{(k)})$ afin de préparer le terrain à l'analyse de l'algorithme de Newton dans le chapitre suivant. Nous supposerons donc que la direction de descente est définie de la manière suivante, où $B^{(k)}$ est une matrice symétrique définie positive

$$d^{(k)} = -B^{(k)}\nabla f(x^{(k)}) \tag{4.1}$$

Le fait que $B^{(k)}$ est définie positive garantit que $\langle d^{(k)}|\nabla f(x^{(k)})\rangle < 0$, c'est-à-dire que $d^{(k)}$ est bien une direction de descente.

Remarque 10. Une méthode de descente où la direction $d^{(k)}$ est définie par la relation (4.1) est souvent appelé *descente de gradient préconditionné*, et la matrice $B^{(k)}$ est appelée *préconditionneur*.

4.1 Choix du pas par rebroussement

4.1.1 Rebroussement naïf

Comme nous souhaitons construire une méthode de descente, il serait assez naturel de considérer le pas suivant,

$$t^{(k)} = \max\{t \geq 0 \mid \exists \ell \in \mathbb{N}, t = 2^{-\ell} \text{ et } f(x^{(k)} + td^{(k)}) \leq f(x^{(k)})\}.$$

Autrement dit étant donné $x^{(k)}, d^{(k)}$ on teste les pas de $1, \frac{1}{2}, \frac{1}{4}, \dots, \frac{1}{2^\ell}$ et on s'arrête dès que que la condition de descente $f(x^{(k)} + 2^{-\ell}d^{(k)}) \leq f(x^{(k)})$ est satisfaite. Cet algorithme, assez naturel, est en fait non-convergent en général. Considérons $f(x) = x^2$ sur $\mathbb{R}$ et $x^{(0)} = 1$, $d^{(k)} = -f'(x^{(k)}) = -2x^{(k)}$. Alors le pas $t^{(k)} = 1$ est admissible et on obtient donc $x^{(k)} = (-1)^k$ qui ne converge pas vers le minimum de f.

4.1.2 Rebroussement d'Armijo

L'idée est de renforcer la condition de descente $f(x^{(k)} + t^{(k)} d^{(k)}) \le f(x^{(k)})$ par la condition plus forte (4.2) appellée "condition d'Armijo" :

Lemme 22. *Soit $f \in \mathcal{C}^1(\mathbb{R}^d)$, $v \in \mathbb{R}^d$ tel que $\langle \nabla f(x) | v \rangle < 0$, et $0 < \alpha < \frac{1}{2}$. Alors, il existe $t_0 > 0$ tel que $\forall t \in [0, t_0]$,*

$$f(x + tv) \le f(x) + \alpha t \langle \nabla f(x) | v \rangle. \tag{4.2}$$

Démonstration. C'est une conséquence immédiate de l'égalité

$$f(x + tv) = f(x) + \alpha t \langle \nabla f(x) | v \rangle + (1 - \alpha) t \langle \nabla f(x) | v \rangle + \mathrm{o}(t).$$

□

Définition 13 (Pas d'Armijo)**.** Soit $f \in \mathcal{C}^1(\mathbb{R}^d)$. Étant donné $x^{(k)}$, $d^{(k)}$ on appellera *pas d'Armijo* de paramètres $0 < \alpha < \frac{1}{2}$ et $0 < \beta < 1$ le pas $t^{(k)}$ défini par

$$\begin{aligned} t^{(k)} = \max\{t \,| \ell \in \mathbb{N}, t = \beta^\ell, \\ f(x^{(k)} + t d^{(k)}) \le f(x^{(k)}) + t\alpha \langle \nabla f(x^{(k)}) | d^{(k)} \rangle\} \end{aligned} \tag{4.3}$$

Remarque 11. En d'autres termes, le pas d'Armijo peut être calculé par la procédure

`pas_armijo`$(x^{(k)}, d^{(k)})$:
  $t \leftarrow 1$
  tant que $f(x^{(k)} + t d^{(k)}) > f(x^{(k)}) + t\alpha \langle \nabla f(x^{(k)}) | d^{(k)} \rangle$:
    $t \leftarrow \beta t$
  retourner t

Définition 14 (Méthode de descente avec pas d'Armijo)**.** On appelle algorithme de descente de gradient préconditionné la méthode itérative suivante, où pour tout $k \in \mathbb{N}$, $B^{(k)} = A^{(k)} A^{(k)}$ et $A^{(k)}$ est une matrices symétriques définies positives :

$$\begin{cases} d^{(k)} = -B^{(k)} \nabla f(x^{(k)}) \\ t^{(k)} = \text{ donné par (4.3)} \\ x^{(k)} = x^{(k)} + t^{(k)} d^{(k)}. \end{cases}$$

4.2 Convergence de l'algorithme de descente de gradient préconditionné à rebroussement

Théorème 23. *Soit $\Omega \subseteq \mathbb{R}^d$ un ouvert convexe et $f \in \mathcal{C}^2(\Omega)$ vérifiant*

(i) le sous-niveau $S = \{x \in \Omega \mid f(x) \leq f(x^{(0)})\}$ est compact.

(ii) $\exists 0 < \lambda < \Lambda,\ \forall x \in S, \forall k \in \mathbb{N},\ \lambda \mathrm{Id} \preceq A^{(k)} \mathrm{D}^2 f(x) A^{(k)} \preceq \Lambda \mathrm{Id}$,

Alors les itérées $x^{(k)}$ de l'algorithme de descente avec pas d'Armijo (Déf. 14) convergent vers l'unique minimiseur global de f sur Ω, et de plus, en posant $c = 2\alpha\lambda \min(1, \beta/\Lambda)$,

$$f(x^{(k+1)}) - f(x^*) \leq (1-c)\left(f(x^{(k)}) - f(x^*)\right). \tag{4.4}$$

Remarque 12. Cette analyse suggère une manière simple d'améliorer l'algorithme de descente de gradient : il suffit de choisir $B^{(k)}$ tel que le conditionnement Λ/λ soit aussi proche de 1 que possible. Dans le cas $f(x_1, x_2) = Kx_1^2 + x_2^2$, dont la hessienne est $H = \mathrm{diag}(K, 1)$ (constante), l'idéal est bien sûr de prendre $B = A^2$ où $A = \mathrm{diag}(K^{-1/2}, 1)$, soit $B = H^{-1}$. Plus généralement, le choix $B^{(k)} = \mathrm{D}^2 f(x^{(k)})^{-1}$ est souvent judicieux : il donne ce qu'on appelle la "méthode de Newton".

Lemme 24. *Soit $f \in \mathcal{C}^2(\Omega)$ et $B^{(k)} = (A^{(k)})^2$ telles que que*

$$\forall x \in S,\ A^{(k)} \mathrm{D}^2 f(x) A^{(k)} \preceq \Lambda \mathrm{Id}.$$

Alors, le pas d'Armijo défini par (4.3) vérifie

$$t^{(k)} \geq \min\left(1, \frac{\beta}{\Lambda}\right).$$

Remarque 13. Cette inégalité permet de dire que l'algorithme `pas_armijo` calculant le pas d'Armijo termine dès que $\beta^k \leq \beta/\Lambda$. En d'autres termes, le nombre d'itération de l'algorithme de recherche du pas est au plus $\log(\Lambda/\beta)/\log(1/\beta)$.

Démonstration. Étant donné $t \in \mathbb{R}$, on pose $x_t = x^{(k)} + td^{(k)}$. Si t est tel que $x_t \in S$, alors le segment $[x_0, x_t]$ est inclus dans S et on montre donc comme dans la preuve du théorème 15 que

$$f(x_t) = f(x^{(k)}) + t\langle \nabla f(x^{(k)}) | d^{(k)} \rangle + \frac{t^2}{2} \langle D^2 f(x_s) d^{(k)} | d^{(k)} \rangle$$

Par (4.1), on a $d^{(k)} = -B^{(k)}\nabla f(x^{(k)}) = -A^2\nabla f(x^{(k)})$ où on a noté $A = A^{(k)}$. Ainsi,

$$\begin{aligned}
\langle D^2 f(x_s)d^{(k)}|d^{(k)}\rangle &= \langle D^2 f(x_s)A^2\nabla f(x^{(k)})|A^2\nabla f(x^{(k)})\rangle \\
&= \langle AD^2 f(x_s)A(A\nabla f(x^{(k)}))|A\nabla f(x^{(k)})\rangle \\
&\leq \Lambda \left\|A\nabla f(x^{(k)})\right\|^2 = \Lambda\langle A^2\nabla f(x^{(k)})|\nabla f(x^{(k)})\rangle = -\langle \nabla f(x^{(k)})|d^{(k)}\rangle
\end{aligned}$$

de sorte que

$$f(x_t) \leq f(x^{(k)}) + (1 - \frac{\Lambda}{2}t)t\langle \nabla f(x^{(k)})|d^{(k)}\rangle.$$

Ainsi, la condition d'Armijo

$$f(x_t) \leq f(x^{(k)}) + \alpha t\langle \nabla f(x^{(k)})|d^{(k)}\rangle$$

est vérifiée dès que $1 - \frac{\Lambda}{2}t \geq \alpha$. Comme $\alpha \leq \frac{1}{2}$, il suffit pour cela que $t \leq \Lambda$. □

Lemme 25. *Soit $X \subseteq \Omega \subseteq \mathbb{R}^d$ avec X convexe et Ω ouvert. Soit $f \in \mathcal{C}^2(\Omega)$ et A une matrice symétrique définie positive vérifiant $\forall x \in X, \mathrm{AD}^2 f(x)A \geq \lambda\mathrm{Id}$. Alors*

$$2\lambda(f(x) - f(x^*)) \leq \|A\nabla f(x)\|^2 .$$

Démonstration. On considère $g(y) = f(Ay)$. Alors $\nabla g(y) = A\nabla f(Ay)$ et $\mathrm{D}^2 g(y) = \mathrm{AD}^2 f(Ay)A$. On a supposé $\mathrm{AD}^2 f(x)A \geq \lambda\mathrm{Id}$, et par le corollaire (18) on a

$$2\lambda(g(y) - g(y^*)) \leq \|\nabla g(y)\|^2 ,$$

où $y^* = A^{-1}x^*$. Ainsi, en posant $y = A^{-1}x$, $2\lambda(f(x) - f(x^*)) \leq \|A\nabla f(x)\|^2$. □

Démonstration du théorème 23. Par définition du pas d'Armijo, on sait que

$$\begin{aligned}
f(x^{(k+1)}) &\leq f(x^{(k)}) + \alpha t^{(k)}\langle \nabla f(x^{(k)})|d^{(k)}\rangle \\
&= f(x^{(k)}) - \alpha t^{(k)}\langle \nabla f(x^{(k)})|B^{(k)}\nabla f(x^{(k)})\rangle \\
&= f(x^{(k)}) - \alpha t^{(k)}\langle A^{(k)}\nabla f(x^{(k)})|A^{(k)}\nabla f(x^{(k)})\rangle \\
&\leq f(x^{(k)}) - \varepsilon \left\|A^{(k)}\nabla f(x^{(k)})\right\|^2 \quad \text{où } \varepsilon = \alpha \min(1, \beta/\Lambda)
\end{aligned}$$

où on a utilisé le lemme 24 pour minorer $t^{(k)}$. En combinant avec le lemme précédent on obtient

$$\begin{aligned}
&f(x^{(k+1)}) \leq f(x^{(k)}) - 2\lambda\varepsilon(f(x^k) - f(x^*)) \\
&f(x^{(k+1)}) - f(x^*) \leq (1 - c)(f(x^{(k)}) - f(x^*)) \text{ où } c = 2\lambda\varepsilon.
\end{aligned}$$

□

Chapitre 5

Méthode de Newton

5.1 Méthode de Newton pure

5.1.1 Construction des itérées

Soit $f \in \mathcal{C}^2(\mathbb{R}^d)$, vérifiant $\forall x \in \mathbb{R}^d, \mathrm{D}^2 f(x) \succ 0$. La méthode de Newton est une méthode itérative, qui repose sur le développement de Taylor à l'ordre 2 de la fonction au point courant $x^{(k)}$: $f(x^{(k)} + d) = g(d) + \mathrm{o}(\|d\|^2)$ où

$$g(d) = f(x^{(k)}) + \langle d | \nabla f(x^{(k)}) \rangle + \frac{1}{2} \langle \mathrm{D}^2 f(x^{(k)}) d | d \rangle.$$

Comme par hypothèse $\mathrm{D}^2 f(x^{(k)}) \succ 0$, la fonction g est strictement convexe et admet un unique minimum, que l'on notera $d^{(k)}$. Ce minimum vérifie

$$\nabla g(d^{(k)}) = \mathrm{D}^2 f(x^{(k)}) d^{(k)} + \nabla f(x^{(k)}) = 0,$$

soit $d^{(k)} = -[\mathrm{D}^2 f(x^{(k)})]^{-1} \nabla f(x^{(k)})$. On arrive à la définition suivante :

Définition 15. Les itérées de la méthode de Newton pure sont données par

$$\begin{cases} d^{(k)} = -[\mathrm{D}^2 f(x^{(k)})]^{-1} \nabla f(x^{(k)}) \\ t^{(k)} = 1 \\ x^{(k+1)} = x^{(k)} + t^{(k)} d^{(k)} \end{cases}$$

Le terme "pur" fait référence au choix du pas $t^{(k)} = 1$, par opposition à la méthode de Newton amortie, introduite dans §5.2.

Remarque 14. La direction $d^{(k)}$ est appelée *direction de Newton.* Il s'agit d'une direction de descente si car $\langle \nabla f(x^{(k)}) | d^{(k)} \rangle = -\langle \nabla f(x^{(k)}) | \mathrm{D}^2 f(x^{(k)}) \nabla f(x^{(k)}) \rangle > 0$, mais en général il est possible que $f(x^{(k+1)}) \not\leq f(x^{(k)})$. La méthode de Newton pure n'est donc pas une méthode de descente.

Remarque 15. Dans le cas $d = 1$. Posons $h(x) = f'(x)$. La méthode de Newton s'écrit alors sous la forme

$$x^{(k+1)} = x^{(k)} - f'(x^{(k)})/f''(x^{(k)}) = x^{(k)} - h(x^{(k)})/h'(x^{(k)}).$$

On reconnaît alors la méthode de Newton "classique" permettant de trouver un zéro de la fonction $h = f'$, ce qui revient dans ce cas à trouver un minimum de la fonction convexe f.

5.1.2 Convergence quadratique locale

Définition 16 (Convergence quadratique)**.** Soit $(u_k)_{k\geq 0}$ une suite de limite u^*. La suite (u_k) converge *quadratiquement* vers u^* s'il existe $\gamma > 0$ telle que

$$\|u_{k+1} - u^*\| \leq \gamma \|u_k - u^*\|^2 .$$

Théorème 26. *Soit $f \in \mathcal{C}^3(\mathbb{R}^d)$ vérifiant*
(i) f admet un minimiseur x^* sur $\mathbb{R}$;
(ii) $\forall x \in \mathbb{R}^d, \mathrm{D}^2 f(x) \succ 0$.
Alors, il existe $r > 0$ tel que pour tout $x^{(0)} \in \mathrm{B}(x^*, r)$, la suite $(x^{(k)})$ construite par la méthode de Newton pure est définie pour tout $k \geq 0$ et converge quadratiquement vers x^*.

Remarque 16. Soit $\varepsilon_k = \gamma \left\|x^{(k)} - x^*\right\|$, où γ est la constante dans la définition de convergence quadratique. Alors, $\varepsilon_{k+1} \leq \varepsilon_k^2$, de sorte que $\varepsilon_k \leq \varepsilon_0^{2^k}$. Ainsi, si l'on souhaite une erreur $\varepsilon_k \leq \eta = 10^{-15}$, il suffit que $\varepsilon_0^{2^k} \leq \eta$ soit $2^k \log_2(\varepsilon_0) \leq \log_2(\eta)$. Ainsi, si $\varepsilon = 1/2$ et en prenant la minoration $\eta \geq 2^{-50}$, il suffit que $2^k \geq 50$, soit $k = 6$.

Démonstration. Soit x^* l'unique minimiseur de f sur $\mathbb{R}^d$ et $R > 0$. Comme f est de classe $\mathcal{C}^3$, il existe une constante L telle que $\forall x, x' \in K \cap \mathrm{B}(x^*, R)$, $\left\|\mathrm{D}^2 f(x) - \mathrm{D}^2 f(x')\right\| \leq L \|x - x'\|$. Soit $x \in K$ et $x_t = (1-t)x^* + tx = x^* + t(x - x^*)$. Alors,

$$\begin{aligned}
\nabla f(x) &= \nabla f(x^*) + \int_0^1 \mathrm{D}^2 f(x_t)(x - x^*)\mathrm{d}t \\
&= \mathrm{D}^2 f(x)(x - x^*) + \int_0^1 (\mathrm{D}^2 f(x_t) - \mathrm{D}^2 f(x))(x - x^*)\mathrm{d}t \\
&= \mathrm{D}^2 f(x)(x - x^*) + R(x, x^*)
\end{aligned}$$

où

$$\|R(x, x^*)\| \leq L \|x - x^*\| \int_0^1 \|x_t - x\| \,\mathrm{d}t = \frac{L}{2} \|x - x^*\|^2$$

On considère maintenant l'application $N : x \in \mathbb{R}^d \mapsto x - \mathrm{D}^2 f(x)^{-1}\nabla f(x)$. Les itérées de l'algorithme de Newton vérifient $x^{(k+1)} = N(x^{(k)})$. On a de plus pour tout $x \in K$,

$$\|N(x) - x^*\| = \left\|x - \mathrm{D}^2 f(x)^{-1}\nabla f(x) - x^*\right\| \leq \left\|\mathrm{D}^2 f(x)^{-1}\right\| \|R(x, x^*)\|^2 \leq \frac{L}{2m} \|x - x^*\|^2 .$$

Pour que l'algorithme de Newton soit bien défini, on chercher $0 < r \leq R$ telle que $N(\mathrm{B}(x^*, r)) \subseteq \mathrm{B}(x^*, \min(r, R))$. Pour cela, il suffit que $\frac{L}{2m}r^2 \leq \min(r, R)$, et on peut donc prendre $r < \min(2m/L, \sqrt{2mR/L})$.

Si $x^{(0)} \in \mathrm{B}(x^*, r)$, on a alors $x^{(1)} = N(x^{(0)}) \in \mathrm{B}(x^*, r)$, et par récurrence la suite des itérées $x^{(k)}$ est bien définie pour tout $k \in \mathbb{N}$. De plus, on a bien

$$\left\|x^{(k+1)} - x^*\right\| = \left\|N(x^{(k)}) - x^*\right\| \leq \frac{L}{2m} \left\|x^{(k)} - x^*\right\|^2 .$$

Soit $\varepsilon_k = \frac{L}{2m} \left\|x^{(k)} - x^*\right\|$, de sorte que $\varepsilon_{k+1} \leq \varepsilon_k^2$ soit par récurrence $\varepsilon_k \leq \varepsilon_0^{2^k}$. De plus, comme $r < 2m/L$, on a $\varepsilon_0 < 1$, de sorte que $\lim_{k\to+\infty} \varepsilon_k = 0$. Ceci prouve la convergence quadratique de la suite $x^{(k)}$ vers x^*. □

5.1.3 Invariance par reparamétrisation affine

Un autre avantage de la méthode de Newton que contrairement à la méthode de descente de gradient, elle est invariante par reparamétrisation affine (c'est-à-dire qu'elle ne dépend pas de la base choisie pour l'espace).

Proposition 27. *Soit $f : \mathbb{R}^d \to \mathbb{R}$, A une matrice carrée inversible et $g : A^{-1}\Omega \to \mathbb{R}^d$ définie par $g(y) = f(Ay)$. Soient maintenant*

$$x^{(k+1)} = x^{(k)} - \mathrm{D}^2 f(x^{(k)})^{-1}\nabla f(x^{(k)})$$

$$y^{(k+1)} = y^{(k)} - \mathrm{D}^2 g(y^{(k)})^{-1}\nabla g(y^{(k)})$$

$$\tilde{x}^{(k)} = Ay^{(k)}.$$

Si l'on suppose de plus que $x^{(0)} = \tilde{x}^{(0)}$, alors $\forall k \in \mathbb{N}$, $\tilde{x}^{(k)} = x^{(k)}$.

Démonstration. Calculons d'abord le gradient et la hessienne de g, en les identifiants dans le développement de Taylor à l'ordre 2 :

$$\begin{aligned} g(y + v) = f(A(y + v)) &= f(Ay) + \langle\nabla f(Ay)|Av\rangle + \frac{1}{2}\langle\mathrm{D}^2 f(Ay)Av|Av\rangle + \mathrm{o}(\|Av\|^2) \\ &= g(y) + \langle A^T\nabla f(Ay)|v\rangle + \frac{1}{2}\langle A^T\mathrm{D}^2 f(Ay)Av|v\rangle + \mathrm{o}(\|v\|^2) \end{aligned}$$

On trouve donc $\nabla g(y) = A^T\nabla f(Ay)$ et $\mathrm{D}^2 g(y) = A^T\mathrm{D}^2 f(Ay)A$, ce qui donne

$$\begin{aligned} y^{(k+1)} &= y^{(k)} - \mathrm{D}^2 g(y^{(k)})^{-1}\nabla g(y^{(k)}) \\ &= y^{(k)} - (A^T\mathrm{D}^2 f(Ay^{(k)})A)^{-1}A^T\nabla f(Ay^{(k)}) \\ &= y^{(k)} - A^{-1}\mathrm{D}^2 f(Ay^{(k)})^{-1}\nabla f(Ay^{(k)}) \end{aligned}$$

Ainsi, en multipliant cette égalité par A on obtient

$$\tilde{x}^{(k+1)} = \tilde{x}^{(k)} - \mathrm{D}^2 f(\tilde{x}^{(k)})^{-1} \nabla f(\tilde{x}^{(k)}).$$

$\square$

5.1.4 Non-convergence globale

Dans ce paragraphe, nous construisons un exemple explicite de fonction pour laquelle la méthode de Newton “pure” ne converge pas pour tout $x^{(0)}$. Considérons $f : x \in \mathbb{R} \mapsto \sqrt{x^2+1}$. Alors,

$$f'(x) = \frac{x}{\sqrt{1+x^2}} \qquad f''(x) = \frac{1}{(\sqrt{x^2+1})^{3/2}},$$

de sorte que f est convexe et même $\frac{1}{\sqrt{r^2+1}^{3/2}}$-fortement convexe sur l'intervalle $[-r, r]$. L'unique minimiseur de f sur $\mathbb{R}$ est $x^* = 0$.

Calculons maintenant les itérées de la méthode de Newton : $d^{(k)}$ est définie par l'équation

$$f''(x^{(k)}) d^{(k)} = -f'(x^{(k)}),$$
$$d^{(k)} = x^{(k)}((x^{(k)})^2 + 1)$$

Ainsi, l'itérée $x^{(k+1)}$ est définie par

$$x^{(k+1)} = x^{(k)} + d^{(k)} = (x^{(k)})^3$$

La suite $x^{(k)}$ définie par cette relation peut avoir trois comportements. Si $\left|x^{(0)}\right| < 1$, alors la suite $x^{(k)}$ converge vers $0 = x^*$ (avec une vitesse cubique !). Si $\left|x^{(0)}\right| > 0$, alors la suite $(\left|x^{(k)}\right|)$ tend vers $+\infty$, là encore très vite. Si $\left|x^{(k)}\right| = 1$, la suite est stationnaire en $1 \neq x^*$ (si $x^{(0)} = 1$) ou alterne entre les deux valeurs ± 1.

5.2 Méthode de Newton amortie

Nous allons voir dans cette partie qu'une modification simple la méthode de Newton pure permet de la rendre globalement convergence.

Définition 17. On appelle algorithme Newton amortie la méthode itérative

$$\begin{cases} d^{(k)} = -\mathrm{D}^2 f(x^{(k)})^{-1} \nabla f(x^{(k)}) \\ t^{(k)} = \mathrm{pas\_armijo}(x^{(k)}, d^{(k)}) \\ x^{(k)} = x^{(k)} + t^{(k)} d^{(k)}. \end{cases},$$

où pas_armijo a été défini dans le chapitre précédent.

La méthode de Newton amortie est un cas particulier d'algorithme de descente de gradient préconditionné, où l'on a choisi comme préconditionneur $B^{(k)} = \mathrm{D}^2 f(x^{(k)})$.

Théorème 28. *Soit $f \in \mathcal{C}^3(\mathbb{R}^d)$ vérifiant*

(i) le sous-niveau $S = \{x \in \mathbb{R}^d \mid f(x) \leq f(x^{(0)})\}$ est compact.

(ii) $\forall x \in S,\ \mathrm{D}^2 f(x) \succ 0$.

Alors les itérées $x^{(k)}$ de la méthode de Newton amortie convergent vers l'unique minimiseur global de x^* de f sur $\mathbb{R}^d$. En outre, il existe $k_0 \in \mathbb{N}$ et une constante $\gamma > 0$ tel que $\gamma \left\|x^{k_0} - x^*\right\| < 1$ et

$$\forall k \geq k_0, \quad \left\|x^{k+1} - x^*\right\| \leq \gamma \left\|x^k - x^*\right\|^2.$$

Remarque 17. Un intérêt pratique de cet algorithme est que le choix de $t^{(k)}$ par rebroussement d'Armijo permet de "automatiquement" de passer d'un régime où la convergence est linéaire ($k \leq k_0$), et expliquée par l'analyse de la méthode de gradient préconditionnée et un second régime ($k \geq k_0$) où on observe une convergence quadratique.

La preuve du théorème se fait en deux étapes. Le lemme suivant permet de vérifier les hypothèses du théorème (23) sur la convergence des méthodes de gradient préconditionné. On obtient alors que $\lim_{k\to+\infty} x^{(k)} = x^*$.

Lemme 29. *Pour tout $k \in \mathbb{N}$, il existe une matrice symétrique $A^{(k)} \succ 0$ telle que $(A^{(k)})^2 = B^{(k)}$. De plus, il existe $0 < \lambda < \Lambda$ tels que*

$$\forall k \in \mathbb{N}, \forall x \in S, \qquad \lambda \mathrm{Id} \leq A^{(k)} \mathrm{D}^2 f(x) A^{(k)} \preceq \Lambda \mathrm{Id}$$

Démonstration. Étant donné $k \in \mathbb{N}$, on définit $B^{(k)} = \mathrm{D}^2 f(x) \succ 0$. Il existe donc une matrice orthogonale P et une matrice diagonale $D = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$ telle que $B^{(k)} = P^T D P$. La matrice $A^{(k)} = P^T \mathrm{diag}(\lambda_1^{1/2}, \dots, \lambda_d^{1/2}) P$ est symétrique définie positive et vérifie $(A^{(k)})^2 = B^{(k)}$. De plus,

$$\langle A^{(k)} \mathrm{D}^2 f(x) A^{(k)} v | v \rangle = \langle \mathrm{D}^2 f(x) A^{(k)} v | A^{(k)} v \rangle.$$

Or, d'après la proposition 20, nous savons qu'il existe $0 < m < M$ tels que $\forall x \in S, m\mathrm{Id} \preceq \mathrm{D}^2 f(x) \preceq M\mathrm{Id}$, de sorte que

$$m \left\|A^{(k)} v\right\|^2 \leq \langle A^{(k)} \mathrm{D}^2 f(x) A^{(k)} v | v \rangle \leq M \left\|A^{(k)} v\right\|^2.$$

Or,

$$\left\|A^{(k)} v\right\|^2 = \langle A^{(k)} v | A^{(k)} v \rangle = \langle (A^{(k)})^2 v | v \rangle = \langle \mathrm{D}^2 f(x^{(k)}) v | v \rangle,$$

d'où

$$m^2 \left\|v\right\|^2 \leq \langle A^{(k)} \mathrm{D}^2 f(x) A^{(k)} v | v \rangle \leq M \left\|v\right\|^2.$$

et on peut donc prendre $\lambda = m^2$, $\Lambda = M^2$. $\square$

On montre ensuite que pour $k \gg 1$, on a nécessairement $t^{(k)} = 1$, de sorte que les itérées de la méthode de Newton amortie coïncident alors avec celles de Newton pure pour $k \gg 1$. Par le théorème 26, nous aurons donc démontré le résultat de convergence quadratique.

Lemme 30. *Il existe* $k_0 \in \mathbb{N}$ *tel que* $\forall k \geq k_0, t^{(k)} = 1$.

Démonstration. Admis. □

Chapitre 6

Projection et gradient projeté

On s'intéresse désormais à des problèmes d'optimisation sous contraintes :

$$\min_{x \in K} f(x) \tag{6.1}$$

où $K \subseteq \mathbb{R}^d$ est un ensemble convexe fermé et $f : \mathbb{R}^d \to \mathbb{R}$ est convexe.

Dans le cas de l'optimisation sous contrainte (K fermé),

$$x^* \in \arg\min_K f \not\Longrightarrow \nabla f(x^*) = 0.$$

Exemple 5. $K = [1, 2] \subseteq \mathbb{R}$ et $f(x) = x^2$. Le minimum de f sur $[1, 2]$ est atteint au point $x^* = 1$, mais évidemment $f'(x^*) = 2 \neq 0$.

Pour pouvoir écrire des algorithmes d'optimisation, il faut d'abord comprendre la conditions d'optimalité.

Proposition 31 (Existence). *Le problème d'optimisation sous contraintes (6.1) admet une solution si une des conditions suivantes est vérifiée :*
(i) K est compact, non vide, et f est continue ;
(ii) K est fermé, non vide, f est continue et $\lim_{\|x\| \to +\infty} f(x) = +\infty$;

Proposition 32 (Unicité). *Si la fonction f est strictement convexe et si l'ensemble K est convexe, alors le problème d'optimisation sous contraintes (6.1) admet au plus une solution.*

6.1 Projection sur un convexe fermé

Théorème 33 (Caractérisation de la projection)**.** *Soit $K \subseteq \mathbb{R}^d$ convexe fermé, non vide. Alors, pour tout point $x \in \mathbb{R}^d$, le problème de minimisation*

$$\min_{q\in K} \|q - x\|^2 \tag{6.2}$$

admet un unique minimiseur p. De plus, p est caractérisé par la condition

$$p \in K \text{ et } \forall q \in K, \langle x - p|p - q\rangle \geq 0. \tag{6.3}$$

Définition 18 (Projection)**.** Soit $K \subseteq \mathbb{R}^d$ convexe fermé non vide et $x \in \mathbb{R}^d$. L'unique minimiseur de (6.2) est appelé *projection de x sur K* et noté $\mathrm{p}_K(x)$.

Démonstration du théorème 33. Soit $f(q) = \|x - q\|^2$, qui est convexe. Montrons que si p est un minimiseur de (6.2), alors p vérifie (6.3). Soit $q \in K$ et $q_t = (1 - t)p + tq$ pour $t \in [0, 1]$. Par convexité, $q_t \in K$. De plus, comme p est un minimiseur de f sur K, on a $f(q_t) \geq f(p)$, ce qui implique

$$\forall 0 < t < 1, \quad \frac{f(q_t) - f(p)}{t} \geq 0 \Longrightarrow \langle \nabla f(p)|q - p\rangle = 2\langle p - x|q - p\rangle \geq 0.$$

Réciproquement, considérons un point $\bar{p}$ vérifiant (6.3). Par convexité de f, on a

$$\forall q \in K, f(q) \geq f(p) + \langle q - p|\nabla f(p)\rangle = f(p) + 2\underbrace{\langle q - p|p - x\rangle}_{\geq 0} \geq f(p).$$

Ainsi, p minimise f sur K et est donc un minimiseur de (6.2). $\square$

Corollaire 34. *Soit $K \subseteq \mathbb{R}^d$ convexe fermé non vide et $x, y \in \mathbb{R}^d$. Alors,*

(i) $\langle x - y|\mathrm{p}_K(x) - \mathrm{p}_K(y)\rangle \geq \|\mathrm{p}_K(x) - \mathrm{p}_K(y)\|^2$

(ii) $\|\mathrm{p}_K(x) - \mathrm{p}_K(y)\| \leq \|x - y\|$

Démonstration. Par le théorème précédent, avec $p = \mathrm{p}_K(x)$ et $q = \mathrm{p}_K(y)$ on a

$$\langle x - \mathrm{p}_K(x)|\mathrm{p}_K(x) - \mathrm{p}_K(y)\rangle \geq 0.$$

En inversant les rôles, on obtient

$$\langle y - \mathrm{p}_K(y)|\mathrm{p}_K(y) - \mathrm{p}_K(x)\rangle \geq 0.$$

En sommant ces deux inégalités on obtient (i). Le point (ii) s'obtient à partir de (i) en utilisant l'inégalité de Cauchy-Schwarz. $\square$

Exemple 6. Soit $e \in \mathbb{R}^d$ non nul et $K = \{\lambda e \mid \lambda \in \mathbb{R}\}$. Alors,

$$\mathrm{p}_K(x) = \frac{\langle x|e\rangle}{\|e\|^2} e.$$

Exemple 7. Soit $K = \{x \in \mathbb{R}^d \mid Ax = b\}$ et $x \in \mathbb{R}^d$. Par (6.3), la projection de x sur K est caractérisée par

$$\begin{aligned} p = p_K(x) \iff p \in K \text{ et } \forall q \in K, \langle x - p|p - q\rangle \geq 0 &\iff p \in K \text{ et } \forall v \in \mathrm{Ker}A, \langle x - p|v\rangle = 0 \\ &\iff p \in K \text{ et } x - p \in (\mathrm{Ker}A)^\perp. \end{aligned}$$

de plus, $(\mathrm{Ker}A)^\perp = \mathrm{Im}A^T$ car

$$\begin{aligned} v \in \mathrm{Ker}A &\iff \forall w, \langle Av|w\rangle = 0 \\ &\iff \forall w, \langle v|A^T w\rangle = 0 \\ &\iff v \in (\mathrm{Im}A^T)^\perp, \end{aligned}$$

d'où l'on déduit $\mathrm{Ker}A = (\mathrm{Im}A^T)^\perp$ soit $(\mathrm{Ker}A)^\perp = \mathrm{Im}A^T$.

Proposition 35. *Si* $\mathrm{Ker}A^T = \{0\}$ *et* $K = \{x \mid Ax = b\}$*, alors*

$$\mathrm{p}_K(x) = x - A^T(AA^T)^{-1}(Ax - b).$$

Démonstration. Par la caractérisation précédente, $p = \mathrm{p}_K(x)$ si et seulement si $p \in K$ et si $x - p \in \mathrm{Im}A^T$, c'est-à-dire s'il existe w tel que $p = x - A^T w$ et $Ap = b$. Le vecteur w est donc caractérisé par

$$A(x - A^T)w = b, \qquad \text{i.e. } AA^T w = Ax - b,$$

soit $w = (AA^T)^{-1}(Ax - b)$. Ceci donne la formule souhaitée pour p. □

Exemple 8. Soit $K = \{q \in \mathbb{R}^d \mid \forall 1 \leq i \leq d, q_i \geq 0\}$.

Si $d = 1$, alors $\mathrm{p}_K(x) = \max(x, 0) =: x^+$. Pour montrer cela, nous utilisons comme précédemment la caractérisation (6.3). Le point $p = x^+ \in K$ est la projection de x sur K si

$$\forall q \in K, \langle x - p|p - q\rangle = (x - x^+)(x^+ - q) \geq 0.$$

Or, si $x \geq 0$, on a $x - x^+ = 0$ de sorte que l'inégalité est vraie. Si $x < 0$, $x - x^+ < 0$ et $x^+ - q < 0$, de sorte que l'inégalité est aussi vraie. De manière générale, nous avons la proposition suivante :

Proposition 36. *Soit* $K = \{q \in \mathbb{R}^d \mid \forall 1 \leq i \leq d, q_i \geq 0\}$ *et* $x \in \mathbb{R}^d$*. Alors,*

$$\mathrm{p}_K(x) = (x_1^+, \ldots, x_d^+).$$

Exemple 9. Soit $K = K_1 \times \ldots \times K_\ell \subseteq \mathbb{R}^d$ où les $K_i \subseteq \mathbb{R}^{d_i}$ sont des convexes fermés non vides et où $d = d_1 + \ldots + d_\ell$. Alors,

$$\mathrm{p}_K(\underbrace{x_1}_{\in\mathbb{R}^{d_1}}, \ldots, \underbrace{x_\ell}_{\in\mathbb{R}^{d_\ell}}) = (\mathrm{p}_{K_1}(x_1), \ldots, \mathrm{p}_{K_\ell}(x_\ell))$$

Exemple 10. Soit $K = \{x \in \mathbb{R}^d \mid \|x\| \leq 1\}$ où $\|\cdot\|$ est la norme euclidienne. Alors,

$$\mathrm{p}_K(x) \begin{cases} x & \text{si } \|x\| \leq 1 \\ \frac{x}{\|x\|} & \text{sinon} \end{cases}.$$

Le cas où $x \in K$ est évident ; supposons donc que $x \notin K$ et posons $p := x/\|x\|$. Alors,

$$\forall q \in K, \langle x - p | p - q\rangle = \left(\frac{1}{\|x\|} - 1\right) \langle x | p - q\rangle$$

Or, par Cauchy-Scharwz et en utilisant $\langle x|p\rangle = \|x\|$, $\langle x|p - q\rangle \geq \|x\| - \|x\| \|q\| \geq 0$. Par la caractérisation (6.3) on obtient comme souhaité $\mathrm{p}_K(x) = p$.

6.2 Condition d'optimalité pour l'optimisation sous contraintes

6.2.1 Condition nécessaire et suffisante d'optimalité

Dans cette partie, nous démontrons plusieurs CNS d'optimalité pour l'optimisation sous contraintes.

À nouveau, x^* peut minimiser une fonction f sur un compact K alors que $\nabla f(x^*) \neq 0$!. Pour écrire des algorithmes d'optimisation corrects, nous devons quelle est la bonne condition nécessaire et suffisante d'optimalité.

Théorème 37. *Soit $K \subseteq \mathbb{R}^d$ un convexe fermé non vide et $f \in \mathcal{C}^1(\mathbb{R}^d)$ une fonction convexe. Alors, les affirmations suivantes sont équivalentes :*
(i) $x^* \in \arg\min_K f$
(ii) $\forall y \in K, \langle x^* - y| - \nabla f(x^*)\rangle \geq 0$ (formulation variationnelle)
(iii) $\exists \tau > 0,\ \forall y \in K, \mathrm{p}_K(x^* - \tau\nabla f(x^*)) = x^*$ (formulation géométrique)
(iv) $\forall \tau > 0,\ \forall y \in K, \mathrm{p}_K(x^* - \tau\nabla f(x^*)) = x^*$

Remarque 18. Ce théorème ne dit rien de l'existence ni ce l'unicité, il permet seulement de caractériser l'optimalité d'un point pour un problème d'optimisation sous contraintes.

Démonstration. (ii) $\implies$ (i) : Par convexité de f, on sait que

$$\forall y \in \mathbb{R}^d,\ f(y) \geq f(x^*) + \langle y - x^* | \nabla f(x^*)\rangle.$$

Or, si $y \in K$, alors $\langle x^* - y | -\nabla f(x^*)\rangle \geq 0$ par hypothèse, de sorte que

$$\forall y \in \mathbb{R}^d,\ \ f(y) \geq f(x^*),$$

i.e. x^* est le minimum de f sur K.

(i) $\Longrightarrow$ (ii) : Soit $y \in K$ et $y_t = (1-t)x^* + ty$ pour $t \in [0,1]$. Par convexité de K, $y_t \in K$ pour tout $t \in [0,1]$. De plus, comme x^* est un minimiseur de f sur K, on a $f(y_t) \geq f(x^*)$, ce qui implique

$$\forall 0 < t < 1, \quad \frac{f(x^* + t(y - x^*)) - f(x^*)}{t} \geq 0.$$

En passant à la limite $t \to 0, t > 0$, on obtient $\langle \nabla f(x^*) | y - x^*\rangle \geq 0$.

(iv) $\Longleftrightarrow$ (iii) $\Longleftrightarrow$ (ii). Soit $t > 0$. Par caractérisation de la projection du point $x = x^* - \tau\nabla f(x^*))$ sur le convexe fermé K, on a

$$\begin{aligned} \mathrm{p}_K(x^* - \tau\nabla f(x^*)) = x^* &\Longleftrightarrow \forall y \in K, \langle (x^* - \tau\nabla f(x^*)) - x^* | x^* - y\rangle \geq 0 \\ &\Longleftrightarrow \forall y \in K, \langle -\nabla f(x^*) | x^* - y\rangle \geq 0, \end{aligned}$$

où on a utilisé $\tau > 0$ pour obtenir la deuxième équivalence. □

Exemple 11. Soit $K = \mathbb{R}^d_+ \subseteq \mathbb{R}^d$. On a démontré précédemment que la projection d'un point x sur K est donnée par

$$\mathrm{p}_K(x) = (\max(x_1, 0), \ldots, \max(x_d, 0)).$$

Soit $f \in \mathcal{C}^1(\mathbb{R}^d)$ une fonction convexe. Combiné au théorème précédent, on obtient

$$\begin{aligned} x^* \in \arg\min_K f &\Longleftrightarrow \forall t > 0,\ \ \mathrm{p}_K(x^* - t\nabla f(x^*)) = x^*, \\ &\Longleftrightarrow \forall t > 0, \forall i, \max\left(x_i^* - t\frac{\partial f}{\partial e_i}(x^*), 0\right) = x_i^* \\ &\Longleftrightarrow \forall i, \forall t > 0, \max\left(x_i^* - t\frac{\partial f}{\partial e_i}(x^*), 0\right) = x_i^* \end{aligned}$$

On vérifie alors que cette propriété est équivalente à

$$\begin{cases} \frac{\partial f}{\partial e_i} = 0 & \text{si } x_i^* > 0 \\ \frac{\partial f}{\partial e_i} \geq 0 & \text{si } x_i^* = 0. \end{cases}$$

6.2.2 Algorithme du gradient projeté

Définition 19. Soit $f \in \mathcal{C}^1(\mathbb{R}^d)$ et $K \subseteq \mathbb{R}^d$ un ensemble convexe fermé non vide. L'algorithme du gradient projeté à pas constant $\tau > 0$ est donné par :

$$x^{(k+1)} = \mathrm{p}_K(x^{(k)} - \tau\nabla f(x^{(k)})) \tag{6.4}$$

Remarque 19. Cette méthode ne mérite le nom d'algorithme que lorsqu'on sait calculer explicitement la projection sur l'ensemble K !

Théorème 38. *On suppose que $f \in \mathcal{C}^2(\mathbb{R}^d)$, et que*

$$\exists 0 < m \leq M \text{ t.q. } \forall x \in \mathbb{R}^d, \quad m\mathrm{Id} \preceq \mathrm{D}^2 f(x) \preceq M\mathrm{Id}. \tag{6.5}$$

Alors la suite définie par (6.4) converge vers l'unique $x^* \in \arg\min_K f$ si le pas τ vérifie

$$0 < \tau < 2m/M^2.$$

On peut écrire $x^{(k+1)} = F(x^{(k)})$ où $F(x) := \mathrm{p}_K(x - \tau\nabla f(x))$. Si la suite $x^{(k)}$ converge vers un point $\bar{x} \in \mathbb{R}^d$, alors en passant à la limite dans l'équation $x^{(k+1)} = F(x^{(k)})$ et en utilisant la continuité de F, on déduit que

$$\bar{x} = F(\bar{x}) = p_K(\bar{x} - \tau\bar{x}),$$

de sorte que par le théorème 37, $\bar{x}$ minimise f sur K. Il reste donc à démontrer que la suite $x^{(k)}$ définie par la récurrence (6.4) converge. Pour cela, nous appliquerons le théorème du point fixe contractant.

Théorème 39 (Point fixe)**.** *Soit $F : \mathbb{R} \to \mathbb{R}^d$ une application contractante, c'est-à-dire qu'il existe $\kappa \in]0, 1[$ tel que $\|F(x) - F(y)\| \leq \kappa \|x - y\|$. Alors :*

(i) F admet un unique point fixe x^* ;

(ii) pour tout $x^{(0)} \in \mathbb{R}^d$, la suite définie par $x^{(k+1)} = F(x^{(k)})$ converge vers x^*.

Nous utiliserons également le lemme suivant.

Lemme 40. *Soit $f \in \mathcal{C}^2(\mathbb{R}^d)$ vérifiant (6.5). Alors, l'application $G : x \mapsto x - \tau\nabla f(x)$ vérifie*

$$\|G(x) - G(y)\|^2 \leq \left(1 - 2\tau m + M^2\tau^2\right) \|x - y\|^2$$

Démonstration. On applique la formule de Taylor avec reste intégral à l'application $h(t) = \nabla f(x_t)$, avec $x_t = x + tv$ et $v = y - x$, ce qui donne

$$h(1) - h(0) = \nabla f(y) - \nabla f(x) = \int_0^1 D^2 f(x_t) v \mathrm{d}t.$$

On en déduit que

$$\begin{aligned}
\|\nabla f(y) - \nabla f(x)\|^2 &= \left\| \int_0^1 D^2 f(x_t) v \mathrm{d}t \right\|^2 \\
&\leq \int_0^1 \left\| D^2 f(x_t) v \right\|^2 \mathrm{d}t \\
&\leq \int_0^1 \left\| D^2 f(x_t) \right\|^2 \|v\|^2 \mathrm{d}t \\
&\leq M^2 \|v\|^2,
\end{aligned}$$

où l'on utilise que pour une matrice symétrique la norme matricielle induite par le produit scalaire est égale au maximum des valeurs absolues des valeurs propres,qui est ici majoré par M. De plus, comme $\mathrm{D}^2 f(x) \succeq m\mathrm{Id}$ pour tout $x \in \mathbb{R}^d$,

$$\begin{aligned}
\|\nabla f(y) - \nabla f(x)\| \, x - y &= \langle \int_0^1 D^2 f(x_t) v \mathrm{d}t | v \rangle \\
&= \int_0^1 \langle D^2 f(x_t) v | v \rangle \mathrm{d}t \\
&\geq m \|v\|^2.
\end{aligned}$$

Ainsi, pour tout $x, y \in \mathbb{R}^d$,

$$\begin{aligned}
\|G(x) - G(y)\|^2 &= \|x - y - \tau(\nabla f(x) - \nabla f(y))\|^2 \\
&= \|x - y\|^2 - 2\tau \langle x - y | \nabla f(x) - \nabla f(y) \rangle + \tau^2 \|\nabla f(x) - \nabla f(y)\|^2 \\
&\leq \left(1 - 2\tau m + M^2 \tau^2\right) \|x - y\|^2
\end{aligned}$$

□

Démonstration du théorème 38. On pose $F(x) = p_K(G(x))$ où $G(x) = x - \tau \nabla f(x)$, de sorte que $x^{(k+1)} = F(x^{(k)})$. En utilisant que la projection p_K est 1-Lipschitzienne, on a

$$\|F(x) - F(y)\| = \|\mathrm{p}_K(G(x)) - \mathrm{p}_K(G(y))\| \leq \|G(x) - G(y)\|.$$

En combinant avec le lemme précédent, on trouve

$$\|F(x) - F(y)\|^2 \leq \|G(x) - G(y)\|^2 \leq \left(1 - 2\tau m + M\tau^2\right) \|x - y\|^2.$$

Cette application est contractante si

$$1 - 2\tau m + M^2 \tau^2 \leq 1 \Longleftrightarrow \tau(M^2 \tau - 2m),$$

ce qui est vrai si $\tau > 0$ et $\tau < 2m/M^2$. Par le théorème du point fixe, on en déduit que dans ce cas l'application F possède un unique point fixe. Par le théorème 37, on sait qu'un point $\bar{x}$ minimise f sur K si et seulement si $F(\bar{x}) = \bar{x}$. On déduit de ce qui précède que le problème de minimisation $\min_K f$ possède une seule solution x^*, qui est l'unique point fixe de F. Par le (ii) du théorème du point fixe, la suite $x^{(k)}$ définie par $x^{(k)} = F(x^{(k)})$ converge vers x^*. □

Chapitre 7

Optimisation avec contraintes d'inégalités

Dans ce chapitre on s'intéresse à des problèmes d'optimisation $\min_K f$ où l'ensemble de contraintes K est défini par des inégalités : étant données des fonctions convexes $c_1, \ldots, c_\ell \in \mathcal{C}^1(\mathbb{R}^d)$, on définit l'ensemble K

$$K = \{x \in \mathbb{R}^d \mid c_1(x) \leq 0, \ldots, c_\ell(x) \leq 0\}.$$

On appelle chacune des fonctions $c_1, \ldots, c_\ell$ une *contrainte* du problème et on note souvent le problème par

$$P := \min_{c_1(x) \leq 0, \ldots, c_\ell(x) \leq 0} f(x)$$

Lemme 41. *Si les fonctions $c_1, \ldots, c_\ell : \mathbb{R}^d \to \mathbb{R}$ sont continues et convexes, alors l'ensemble $K = \{x \in \mathbb{R}^d \mid \forall 1 \leq i \leq \ell,\ c_i(x) \leq 0\}$ est fermé et convexe.*

Démonstration. Montrons d'abord que K est fermé : soit $(x_n)_n$ une suite d'éléments de K convergent vers une limite x. Par hypothèse, comme $x_n \in K$, $c_i(x_n) \leq 0$. En passant à la limite $n \to +\infty$ et en utilisant la continuité de c_i on en déduit que $c_i(x) \leq 0$. Ainsi x appartient à K, et K est donc fermé.

On montre maintenant la convexité de K. Soient $x, y \in K$ et $x_t = (1-t)x + ty$. Comme $x, y \in K$, on a pour tout i, $c_i(x) \leq 0$ et $c_i(y) \leq 0$. Pour $t \in [0,1]$ on a donc

$$c_i(x_t) \leq (1-t)c_i(x) + tc_i(y) \leq 0,$$

de sorte que $x_t \in K$. Ceci démontre que K est convexe. □

Exemple 12 (Simplexe). Considérons l'ensemble K définit par

$$K = \left\{ x \in \mathbb{R}^d \mid \forall i, x_i \geq 0 \text{ et } \sum_{1 \leq i \leq d} x_i = 1 \right\},$$

qui décrit l'ensemble des mesures de probabilités supportés sur $\{1, \ldots, d\}$. On peut écrire cet ensemble sous la forme ci-dessus avec

$$c_1(x) = -x_1, \ldots, c_d(x) = -x_d, c_{d+1}(x) = \sum_{1\leq i\leq d} x_i - 1, c_{d+2}(x) = -(\sum_{1\leq i\leq d} x_i - 1).$$

7.1 Méthode de pénalisation

L'idée de cette méthode est d'approcher le problème d'optimisation avec contraintes $\min_K f$ où $K = \{x \mid \forall i,\ c_i(x) \leq 0\}$ par un problème d'optimisation sans contraintes $\min_{\mathbb{R}^d} f_\varepsilon$ où f_ε est la somme de f et de termes qui “explosent” lorsque $c_i(x) > 0$. Précisément, pour $\varepsilon > 0$ on pose

$$P_\varepsilon := \min_{x\in\mathbb{R}^d} f_\varepsilon(x) \text{ où } f_\varepsilon(x) := f(x) + \frac{1}{\varepsilon}\sum_{i=1}^{\ell} \max(c_i(x), 0)^2.$$

Dans cette formulation du problème, les points $x \in \mathbb{R}^d$ tels que $c_i(x) > 0$ sont *pénalisés* au sens où si ε est très petit, $\frac{1}{\varepsilon}\max(c_i(x), 0)^2$ peut être très grand, ce qui dissuade le choix de ce point dans le problème d'optimisation. Lorsque $\varepsilon \to 0$, les points vérifiants sont “infiniment pénalisés” et deviennent en fait interdits :

$$\forall x \in \mathbb{R}^d, \quad \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\sum_{i=1}^{\ell} \max(c_i(x), 0)^2 = \begin{cases} 0 & \text{si } c_i(x) \leq 0 \\ +\infty & \text{sinon} \end{cases}$$

Cette intuition est précisée par la proposition suivante.

Proposition 42. *Supposons que $f \in \mathcal{C}^0(\mathbb{R}^d)$ est strictement convexe et vérifie* $\lim_{\|x\|\to+\infty} f(x) = +\infty$. *Alors :*

(i) Les problèmes P et P_ε (pour $\varepsilon > 0$) admettent chacun un unique minimiseur, noté x^* et x^*_ε.

(ii) $\lim_{\varepsilon\to 0} x^*_\varepsilon = x^*$.

Démonstration. (i) L'existence d'une solution au problème d'optimisation sans contrainte P_ε se déduit du fait que f_ε est continue et que

$$\lim_{\|x\|\to+\infty} f_\varepsilon(x) \geq \lim_{\|x\|\to+\infty} f(x) = +\infty,$$

tandis que l'existence de solution au problème avec contraintes P se déduit de ce que $\lim_{\|x\|\to+\infty} f(x) = +\infty$ et que K est fermé (Lemme 41). Pour l'unicité de la solution au problème P, il suffit de remarquer que f est strictement convexe (hypothèse) et que K est convexe (Lemme 41). Enfin, pour l'unicité de la solution au problème P_ε il faut montrer que f_ε est strictement convexe. Soient $x, y \in \mathbb{R}^d$, $t \in [0, 1]$ et $x_t = (1 - t)x + ty$. Alors,

$$c_i(x_t) \leq (1 - t)c_i(x) + tc_i(y)$$

de sorte que

$\max(c_i(x_t), 0) \leq \max((1-t)c_i(x)+tc_i(y), 0) \leq (1-t)\max(c_i(x), 0)+t\max(c_i(y), 0).$

Ainsi, en utilisant la convexité de $r \in \mathbb{R} \mapsto r^2$, on a

$$\begin{aligned}\max(c_i(x_t), 0)^2 &\leq [(1-t)\max(c_i(x), 0) + t\max(c_i(y), 0)]^2 \\ &\leq (1-t)\max(c_i(x), 0)^2 + t\max(c_i(y), 0)^2\end{aligned}$$

On en déduit que les fonctions $x \mapsto \max(c_i(x), 0)^2$ sont convexes, et f_ε est donc strictement convexe comme combinaison linéaire à coefficients positifs de fonctions convexes et d'une fonction strictement convexe.

(ii) Soit $x^* \in K$ le minimiseur de P. Alors, comme $c_i(x^*) \leq 0$, on a

$$P_\varepsilon \leq f_\varepsilon(x^*) = f(x^*) + \frac{1}{\varepsilon}\sum_{i=1}^{\ell}\max(c_i(x^*), 0)^2 = f(x^*) = P.$$

Ainsi,

$$P_\varepsilon = f(x_\varepsilon^*) + \frac{1}{\varepsilon}\sum_{1\leq i\leq \ell}\max(c_i(x_\varepsilon^*), 0)^2 \leq P.$$

Comme $\lim_{\|x\|\to+\infty} f(x) = +\infty$, la fonction f est minorée, par exemple $f \geq m \in \mathbb{R}$. On déduit donc que

$$\forall i \in \{1, \ldots, \ell\}, \quad \frac{1}{\varepsilon}\max(c_i(x_\varepsilon^*), 0)^2 \leq \frac{1}{\varepsilon}\sum_{1\leq j\leq \ell}\max(c_j(x_\varepsilon^*), 0)^2 \leq P - f(x_\varepsilon^*) \leq P - m,$$

ou encore

$$\max(c_i(x_\varepsilon^*), 0)^2 \leq \varepsilon(P-m).$$

Soit $\bar{x}$ une valeur d'adhérence de la suite x_ε^* lorsque $\varepsilon \to 0$. Alors, par continuité,

$$\max(c_i(\bar{x}), 0)^2 \leq 0,$$

ce qui montre que $\bar{x} \in K$. De plus, pour tout $\varepsilon > 0$,

$$f(x_\varepsilon) \leq P_\varepsilon \leq P,$$

de sorte que $f(\bar{x}) \leq P$. On en déduit que $\bar{x}$ minimise f sur K, soit $\bar{x} = x^*$. Pour conclure, on remarque que la suite (x_ε^*) est bornée et admet une unique valeur d'adhérence, de sorte que $\lim_{\varepsilon\to 0} x_\varepsilon^* = x^*$. □

Lemme 43. *Soit $p : t \in \mathbb{R} \mapsto \max(t, 0)^2$. Alors $p \in \mathcal{C}^1(\mathbb{R})$ et $p'(t) = 2\max(t, 0)$*

Démonstration. Comme $p(t) = O(t^2)$, p est différentiable en zéro et $p'(0) = 0$. D'autre part, sur $\mathbb{R}_+^*$, $p(t) = t^2$ et $p'(t) = 2t$, et sur $\mathbb{R}_-^*$, $p(t) = 0$, soit $p'(t) = 0$. Ainsi,

$$p'(t) = \begin{cases} 0 & \text{si } t < 0 \\ 0 & \text{si } t = 0 \\ 2t & \text{si } t > 0 \end{cases}$$

Ainsi, $p'(t)$ est continue et $p(t) = 2\max(t, 0)$. □

Proposition 44. *Supposons que $f \in \mathcal{C}^1(\mathbb{R}^d)$ est strictement convexe. Alors,*

$$x_\varepsilon^* \in \arg\min_{\mathbb{R}^d} f_\varepsilon \Longleftrightarrow \nabla f(x_\varepsilon^*) + \frac{2}{\varepsilon}\sum_{i=1}^{\ell} \max(c_i(x_\varepsilon^*), 0)\nabla c_i(x_\varepsilon^*) = 0. \tag{7.1}$$

Démonstration. Nous avons déjà démontré que f_ε est strictement convexe, et $f_\varepsilon \in \mathcal{C}^1(\mathbb{R}^d)$ par le lemme précédente, de sorte que par caractérisation de l'optimalité dans le problème d'optimisation *sans contrainte* P_ε,

$$x_\varepsilon^* \in \arg\min_{\mathbb{R}^d} f_\varepsilon \Longleftrightarrow \nabla f_\varepsilon(x_\varepsilon^*) = 0.$$

D'autre part, en utilisant la fonction p introduite dans le lemme précédent on peut écrire

$$f_\varepsilon(x) = f(x) + \frac{1}{\varepsilon}\sum_{1\leq i\leq \ell} p(c_i(x)),$$

soit

$$\begin{aligned} \nabla f_\varepsilon(x) &= \nabla f(x) + \frac{2}{\varepsilon}\sum_{1\leq i\leq \ell} p'(c_i(x))\nabla c_i(x). \\ &= \nabla f(x) + \frac{2}{\varepsilon}\sum_{i=1}^{\ell} \max(c_i(x), 0)\nabla c_i(x) \end{aligned}$$

On en déduit l'équivalence annoncée. □

7.2 Théorème de Karush-Kush-Tucker

L'objet du reste de ce chapitre est d'énoncer et de démontrer le théorème de Karush-Kush-Tucker. Ce théorème permet de donner une condition nécessaire et suffisante d'optimalité pour les problèmes d'optimisation avec contraintes d'inégalité.

Théorème 45 (Karush-Kush-Tucker)**.** *Soient* $f, c_i : \mathbb{R}^d \to \mathbb{R}$ *vérifiant :*

— $f \in \mathcal{C}^1(\mathbb{R}^d)$ *convexe ;*

— $c_i(x) = \langle a_i | x\rangle - b_i$ *pour* $\forall i \in \{1, \dots, \ell\}$ *;*

et soit

$$K = \{x \in \mathbb{R}^d \mid \forall i \in \{1, \dots, \ell\},\ c_i(x) \leq 0\}.$$

Alors,

$$x^* \in \arg\min_K f \Longleftrightarrow \exists \lambda \in \mathbb{R}^\ell \ t.q. \begin{cases} -\nabla f(x^*) = \sum_{i=1}^\ell \lambda_i \nabla c_i(x^*) & \\ x^* \in K & \\ \lambda_i \geq 0 & \forall i \in \{1, \dots, \ell\} \\ \lambda_i c_i(x^*) = 0 & \forall i \in \{1, \dots, \ell\} \end{cases}$$

Remarque 20. Par analogie avec le théorème des extrémas liés, le vecteur λ apparaissant dans ce théorème est souvent appelé *multiplicateur de Lagrange*.

Remarque 21. Les quatres conditions apparaissant dans ce théorème ont un nom :

— la première condition ($-\nabla f(x^*) = \dots$) est appelée *condition d'équilibre* ;

— la deuxième ($x^* \in K$) est l'*admissibilité du point* x ;

— la troisième ($\lambda_i \geq 0$) est l'*admissibilité du multiplicateur de Lagrange* λ ;

— et la quatrième ($\lambda_i c_i(x)$) est la *condition de complémentarité*.

Il est important de n'oublier aucune de ces quatre conditions lorsqu'on applique le théorème de Karush-Kuhn-Tucker (souvent abrégé KKT).

Nous commençons pas démontrer le sens direct de ce théorème, en utilisant la méthode de pénalisation, et plus précisément en passant à la limite dans l'équation d'optimalité du problème pénalisé (7.1).

Démonstration du sens direct ($\Longrightarrow$) du théorème 47. Soit x^* un minimiseur de f sur K. On commence la démonstration en supposant que f est strictement convexe, et on montrera comment en déduire le cas général. Soit $\varepsilon_N = \frac{1}{N}$ et $x_N^* := x_{\varepsilon_N}^*$ le minimiseur du problème pénalisé P_{ε_N} :

$$x_N^* \in \arg\min_{x \in \mathbb{R}^d} f(x) + \frac{1}{\varepsilon} \sum_{i=1}^{\ell} \max(c_i(x), 0)^2.$$

Par la condition d'optimalité (7.1) pour le problème P_{ε_N}, x_N^* vérifie

$$\nabla f(x_N^*) + \frac{2}{\varepsilon_N} \sum_{i=1}^{\ell} \max(c_i(x_N^*), 0) \nabla c_i(x_N^*) = 0.$$

Soit I l'ensemble des $i \in \{1, \dots, \ell\}$ tels que $c_i(x^*) < 0$. Par la proposition 42, x_N^* converge vers x^*, de sorte qu'il existe N_0 tel que

$$\forall N \geq N_0, \forall i \in I, c_i(x_N^*) < 0.$$

Ainsi, pour $N \geq N_0$, et en utilisant $\nabla c_i(x_N^*) = a_i$ (car $c_i(x) = \langle x | a_i \rangle - b_i$),

$$\begin{aligned} -\nabla f(x_N^*) &= \frac{2}{\varepsilon_N} \sum_{i=1}^{\ell} \max(c_i(x_N^*), 0) \nabla c_i(x_N^*) \\ &= \frac{2}{\varepsilon_N} \sum_{i \notin I} \max(c_i(x_N^*), 0) a_i \\ &\in C \end{aligned}$$

où C est l'ensemble des combinaisons linéaires à coefficients positifs des $(a_i)_{i \notin I}$, que l'on peut écrire

$$C := \left\{ \sum_{1 \leq i \leq N} \lambda_i a_i \mid \lambda \in \mathbb{R}_+^\ell \text{ tq } \forall i \in I, \lambda_i = 0 \right\}.$$

Comme l'ensemble C est fermé (lemme 46), en utilisant la continuité de ∇f ($f \in \mathcal{C}^1(\mathbb{R}^d)$) et $\lim_{N \to +\infty} x_N^* = x_N$, on en déduit que

$$-\nabla f(x^*) \in C.$$

Ainsi, il existe $\lambda \in \mathbb{R}_+^N$ tel que

$$-\nabla f(x^*) = \sum_i \lambda_i a_i = \sum_i \lambda_i \nabla c_i(x^*),$$

qui par définition de l'ensemble C vérifie en outre $\forall i \in I,\ \lambda_i = 0$. Ainsi, si $i \in I$, $c_i(x^*)\lambda_i = 0$, et si $i \notin I$, $c_i(x^*) = 0$ et on a aussi $c_i(x^*)\lambda_i = 0$.

Pour finir, nous avions supposé que la fonction f était strictement convexe. Si ça n'est pas le cas, on remplace f par $\tilde{f}(x) = f(x) + \|x - x^*\|^2$. La fonction $\tilde{f}$ est strictement convexe (somme de convexe et strictement convexe) et a aussi x^* pour minimiseur. On peut donc lui appliquer la démonstration précédente : il existe $\lambda \in \mathbb{R}^\ell$ vérifiant

$$\begin{cases} -\nabla \tilde{f}(x^*) = \sum_{1 \leq i \leq \ell} \lambda_i \nabla c_i(x^*) & \\ x^* \in K & \\ \lambda_i \geq 0 & \forall i \in \{1, \dots, \ell\} \\ \lambda_i c_i(x) = 0 & \forall i \in \{1, \dots, \ell\}. \end{cases}$$

Pour conclure, il suffit de remarquer que $\nabla \tilde{f}(x^*) = \nabla f(x^*)$. □

Lemme 46. *Soit $a_1, \dots, a_\ell \in \mathbb{R}^d$ et $I \subseteq \{1, \dots, \ell\}$. Alors l'ensemble C défini ci-dessous est fermé :*

$$C := \left\{ \sum_{1 \leq i \leq l} \lambda_i a_i \mid \lambda \in \mathbb{R}_+^\ell \right\}.$$

Démonstration. Supposons d'abord que les $(a_i)_{i=1,\cdots,l}$ sont linéament indépendants. Soit $v_n = \sum_{i=1} \lambda_i^n a_i$ une suite d'éléments de C convergeant vers une limite $v \in \mathbb{R}^d$. Alors il est clair que chaque suite (λ_i^n) converge dans $\mathbb{R}_+$ vers une limite $\lambda_i \geq 0$ puisque les vecteurs $(a_i)_{i=1,\cdots,l}$ forment une base de l'espace qu'ils engendrent. On a donc $v = \sum_{i=1}^l \lambda_i a_i \in C$. Si les vecteurs $(a_i)_{i=1,\cdots,l}$ sont linéairement dépendants, il existe une relation de la forme $\sum_{i=1}^l \mu_i a_i = 0$ et on peut supposer que l'un au moins des coefficients μ_i est strictement positif. On peut aussi écrire

$$v = \sum_{i=1}^{l} (\lambda_i + t\mu_i) a_i, \quad t = \min\left(-\frac{\lambda_i}{\mu_i} \right)$$

tel que tous les $\lambda_i + t\mu_i$ sont positifs sauf au moins un qui est nul $\exists i_0 \; \lambda_{i_0} + t\mu_{i_0} = 0$. Ce raisonnement montre que

$$C = \bigcup_{i=1}^{l} \left\{ \sum_{i \neq i_0}^{l} \lambda_i a_i \mid \lambda_i \geq 0, \; \forall i \right\}.$$

Il suffit de répéter l'argument jusqu'au point où le cône C est exprimé comme l'union finie de cônes associés à des vecteurs a_i, qui sont linéairement indépendants et donc fermés, de la manière décrite dans la première partie de l'argument.

□

Démonstration du sens indirect ($\Longleftarrow$) du théorème 47. Soit $x^* \in \mathbb{R}^d$ et $\lambda \in \mathbb{R}^\ell$ vérifiant les quatre conditions du théorèmes. Comme la fonction f est convexe,

$$\begin{aligned} \forall x \in \mathbb{R}^d, \; f(x) &\geq f(x^*) + \langle x - x^* | \nabla f(x^*) \rangle \\ &= f(x^*) - \sum_{i=1}^{\ell} \lambda_i \langle x - x^* | \nabla c_i(x^*) \rangle, \end{aligned} \tag{7.2}$$

où l'on a utilisé la condition d'équilibre ($-\nabla f(x^*) = ...$) pour obtenir l'égalité de la deuxième ligne. Comme la contrainte c_i est convexe (car $c_i(x) = \langle x | a_i \rangle - b_i$), on a

$$\forall x \in \mathbb{R}^d, \; c_i(x) \geq c_i(x^*) + \langle x - x^* | \nabla c_i(x^*) \rangle,$$

soit

$$\forall x \in \mathbb{R}^d, \; -\langle x - x^* | \nabla c_i(x^*) \rangle \geq c_i(x^*) - c_i(x), \tag{7.3}$$

En combinant les inégalités (7.2) et (7.3), et en utilisant à nouveau la condition d'admissibilité de λ ($\lambda_i \geq 0$), on en déduit

$$\forall x \in \mathbb{R}^d,\ f(x) \geq f(x^*) + \sum_{i=1}^{\ell} \lambda_i (c_i(x^*) - c_i(x)).$$

Par la condition de complémentarité ($\lambda_i c_i(x^*) = 0$), on a

$$\forall x \in \mathbb{R}^d,\ f(x) \geq f(x^*) - \sum_{i=1}^{\ell} \lambda_i c_i(x).$$

Si $x \in K$, alors $c_i(x) \leq 0$. En combinant avec la condition d'admissibilité de λ ($\lambda_i \geq 0$), on obtient $\lambda_i c_i(x) \leq 0$, ce qui donne finalement

$$\forall x \in K,\ f(x) \geq f(x^*).$$

Le point x^* est dans K (par la condition d'admissibilité) et $f(x^*) \leq f(x)$ pour tout autre $x \in K$: ceci montre bien que $x^* \in \arg\min_K f$. □

7.2.1 Théorème de Karush-Kush-Tucker : le cas général

Définition 20. Soit $x \in K$. L'ensemble $A(x) = \{i \in \{0, \cdots, l\},\ c_i(x) = 0\}$ est appelé l'ensemble des contraintes actives en x.

Définition 21. On dit que les contraintes $c_i(x) \leq 0$ sont qualifiées en $x \in K$ si et seulement si il existe une direction $v \in \mathbb{R}^d$ telle que l'on ait pour tout $i \in A(x)$

$$\begin{aligned}&\text{ou bien } \langle \nabla c_i(x) | v \rangle < 0\\ &\text{ou bien } \langle \nabla c_i(x) | v \rangle = 0 \text{ et } c_i \text{ est affine.}\end{aligned}$$

On a alors la version suivante du théorème KKT

Théorème 47 (Karush-Kush-Tucker)**.** *Soient* $f, c_i : \mathbb{R}^d \to \mathbb{R}$ *vérifiant :*

— $f \in \mathcal{C}^1(\mathbb{R}^d)$ *convexe ;*

— $c_i(x) \in \mathcal{C}^1(\mathbb{R}^d)$ *convexe pour* $\forall i \in \{1, \ldots, \ell\}$ *;*

et soit

$$K = \{x \in \mathbb{R}^d \mid \forall i \in \{1, \ldots, \ell\},\ c_i(x) \leq 0\}.$$

Soit x^* *un minimiseur de* f *sur* K*. Supposons que les contraintes soient qualifiées en* x^* *au sens de la définition 21. Alors,*

$$x^* \in \arg\min_K f \iff \exists \lambda \in \mathbb{R}^\ell \ t.q. \begin{cases} -\nabla f(x^*) = \sum_{i=1}^{\ell} \lambda_i \nabla c_i(x^*) & \\ x^* \in K & \\ \lambda_i \geq 0 & \forall i \in \{1, \ldots, \ell\} \\ \lambda_i c_i(x^*) = 0 & \forall i \in \{1, \ldots, \ell\} \end{cases}$$

Démonstration. On donnera juste la démonstration du sens direct du théorème. Considérons l'ensemble

$$K(x^*) = \{v \in \mathbb{R}^d \mid \langle \nabla c_i(x^*) | v \rangle \leq 0 \ \forall i \in A(x^*)\}$$

. Soit $\bar{v}$ une direction admissible satisfaisant 21, $v \in K(x^*)$ et un réel $\delta > 0$. Nous allons montrer que $x^* + \varepsilon(v + \delta\bar{v}) \in K$ pour tout réel $\varepsilon > 0$ assez petit. Il faut examiner trois cas

(C1) Si $i \notin A(x^*)$, on $c_i(x^*) < 0$ et $c_i(x^* + \varepsilon(v + \delta\bar{v})) < 0$ par continuité si ε est assez petit ;

(C2) Si $i \in A(x^*)$ et $\langle \nabla c_i(x^*) | \bar{v} \rangle < 0$, alors

$$\begin{aligned} c_i(x_* + \varepsilon(v + \delta\bar{v})) &= c_i(x_*) + \varepsilon \langle \nabla c_i(x_*), v + \delta\bar{v} \rangle + o(\varepsilon) \\ &= 0 + \varepsilon \underbrace{\langle \nabla c_i(x_*), v \rangle}_{\leq 0} + \varepsilon\delta \underbrace{\langle \nabla c_i(x_*), \bar{v} \rangle}_{<0} + o(\varepsilon) \\ &\leq 0 \text{ pour } \varepsilon \text{ assez petit ;} \end{aligned}$$

(C3) Enfin, si $i \in A(x^*)$ et $\langle \nabla c_i(x^*) | \bar{v} \rangle = 0$ alors c_i est affine et

$$\begin{aligned} c_i(x_* + \varepsilon(v + \delta\bar{v})) &= c_i(x_*) + \varepsilon \langle \nabla c_i(x_*), v + \delta\bar{v} \rangle \\ &= \varepsilon \langle \nabla c_i(x_*), v \rangle \leq 0. \end{aligned}$$

Finalement, si x^* est un minimum de f sur K, on a

$$\frac{f(x_* + \varepsilon(v + \delta\bar{v})) - f(x_*)}{\varepsilon} \geq 0$$

et en passant à la limite pour $\varepsilon \to 0$ on obtient

$$\langle \nabla f(x^*) | v + \delta\bar{v} \rangle \geq 0,\ \forall v \in K(x^*),\ \forall \delta > 0.$$

Ceci implique que $\langle \nabla f(x^*) | v \rangle \geq 0 \ \forall v \in K(x^*)$ et on termine la démonstration grâce au Lemme de Farkas ci-dessous. $\square$

Lemme 48 (de Farkas)**.** *Soient* $a_1, \cdots, a_l$ *des éléments fixés de* $\mathbb{R}^d$. *On considère les ensembles*

$$\tilde{C} = \{v \in \mathbb{R}^d \mid \langle a_i | v \rangle \geq 0,\ \forall i\},$$

et

$$C := \left\{ \sum_{1 \leq i \leq N} \lambda_i a_i \mid \lambda \in \mathbb{R}_+^\ell \right\}.$$

Alors pour tout $p \in \mathbb{R}^d$ *on a*

$$\langle p | v \rangle \geq 0,\ \forall v \in \bar{C} \textit{si et seulement si}\, p \in C.$$

Bibliographie

[1] Hedy Attouch, Giuseppe Buttazzo, and Gérard Michaille. *Variational analysis in Sobolev and BV spaces : applications to PDEs and optimization*, volume 17. SIAM, 2014.

[2] Stephen Boyd and Lieven Vandenberghe. *Convex optimization.* Cambridge university press, 2004.

[3] Guillaume Carlier. *Classical and Modern Optimization.* World Scientific, 2022.

[4] Jean Dieudonné. Elements d'analyse. tome ii : Chapitres xii a xv. 1969.

[5] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Fundamentals of convex analysis.* Springer Science & Business Media, 2001.

[6] Robert R. Phelps. *Convex functions, monotone operators and differentiability.* Springer-Verlag, 1989.

I. Pour commencer

Motivation On s'intéresse à la minimisation sur $\Omega = \mathbb{R}^N$ de fonction de la forme

$$f(x) = \frac{1}{2}\langle x|Qx\rangle + \langle b|x\rangle \tag{1}$$

où $Q \in \mathcal{M}_N(\mathbb{R})$ est une matrice symétrique définie positive et $b \in \mathbb{R}^N$ est un vecteur. Ces fonction apparaissent dans de nombreuses applications, et notamment dans la méthode des moindres carrés. On rappelle les définitions suivantes:

(i) Une matrice symétrique Q est dite *positive* (ce qu'on note $Q \succeq 0$) si et seulement $\forall x \in \mathbb{R}^N, \langle x|Qx\rangle \geq 0$.

(ii) Q est dite *définie positive* si $\forall x \in \mathbb{R}^N \setminus \{0\}, \langle x|Qx\rangle > 0$.

Exercice 0. *Exemple explicite*. On considère $Q = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$, $b = (0,0)$.

1. Montrer que si la fonction f définie par (1) est convexe, alors $\lambda_1, \lambda_2 \geq 0$.
2. Montrer que f est minorée si et seulement si $\lambda_1, \lambda_2 \geq 0$.
3. Montrer que $x = (0,0)$ est un point critique de f (i.e. $\nabla f(x^*) = 0$) mais que c'est un minimiseur global si et seulement si $\lambda_1, \lambda_2 \geq 0$.

Exercice 1. *Convexité*. Dans cet premier exercice, nous étudions la convexité de f. On suppose que Q est symétrique, mais pas nécessairement positive.

1. Montrer que $f(x) = \frac{1}{2}\sum_{1\leq i,j\leq N} Q_{ij}x_ix_j + \sum_{1\leq i\leq N} b_ix_i$, $\nabla f(x) = Qx + b$ et $\mathrm{D}^2 f(x) = Q$.
2. En déduire que f est convexe si et seulement si Q est positive (on pourra utiliser la caractérisation de la convexité utilisant ∇f).
3. Démontrer les égalités suivantes (on utilisera la symétrie de Q):

$$\begin{aligned} f((1-t)x+ty) - (1-t)f(x) - tf(y) &= -\frac{t(1-t)}{2}\langle Qx|x\rangle - \frac{t(1-t)}{2}\langle Qy|y\rangle + t(1-t)\langle Qx|y\rangle \\ &= -\frac{t(1-t)}{2}\langle Q(x-y)|x-y\rangle \end{aligned}$$

   En déduire que f est strictement convexe si et seulement si Q est définie positive.
4. Soit Q une matrice définie positive.
   (a) Montrer que la fonction $q_Q : x \mapsto \langle x|Qx\rangle$ atteint son minimum m sur l'ensemble $K = \{x \in \mathbb{R}^N \mid \|x\| = 1\}$, puis que $m > 0$.
   (b) En déduire que $\forall x \in K, \langle x|Qx\rangle \geq m$, puis que $\forall x \in \mathbb{R}^d, \langle x|Qx\rangle \geq m\|x\|^2$.
   (c) Démontrer que la fonction f admet un minimiseur sur $\mathbb{R}^d$.

Dans les exercices 2 et 3, on suppose que Q est symétrique définie positive

Exercice 2. *Caractérisation du minimiseur*. Montrer que f admet un unique minimiseur x^* sur $\mathbb{R}^N$, caractérisé par l'équation $Qx^* + b = 0$.

Exercice 3. *Descente de gradient à pas optimal.* On considère l'algorithme de descente de gradient à pas optimal pour une fonction f. Les itérées $(x^{(k)})_{k\geq 0}$ de cet algorithmes sont définies de manière itérative par $x^{(0)} \in \mathbb{R}^N$ puis:

$$\begin{cases} d^{(k)} = -\nabla f(x^{(k)}) \\ t^{(k)} = \arg\min_{t\in\mathbb{R}} f(x^{(k)} + td^{(k)}) \\ x^{(k+1)} = x^{(k)} + t^{(k)} d^{(k)} \end{cases}$$

La deuxième ligne signifie que $t^{(k)}$ est le minimiseur de $t \mapsto f(x^{(k)} + td^{(k)})$. On s'intéresse au cas $f(x) = \frac{1}{2}\langle Qx|x\rangle + \langle b|x\rangle$.

1. Soit $x \in \mathbb{R}^N$ et $v \in \mathbb{R}^N$ et $g : \mathbb{R} \to \mathbb{R}$ défini par $g(t) = f(x + tv)$.
   (i) Montrer que g est un polynôme de degré 2 dont le coefficient dominant est > 0.
   (ii) Donner l'expression de l'unique minimiseur t^* de $t \in \mathbb{R} \mapsto f(x + tv)$.
2. En déduire que les itérées de la méthode de descente de gradient à pas optimal vérifient

$$\begin{cases} d^{(k)} = -(Qx^{(k)} + b) \\ t^{(k)} = \frac{\langle d^{(k)}|d^{(k)}\rangle}{\langle d^{(k)}|Qd^{(k)}\rangle} \\ x^{(k+1)} = x^{(k)} + t^{(k)} d^{(k)} \end{cases}$$

Exercice 4. *Application aux moindres carrés.* Dans cette question, on s'intéresse au problème de minimisation suivant,

$$\min_{x\in\mathbb{R}^N} \frac{1}{2} \|Ax - y\|^2 ,$$

où A est une matrice de taille $M \times N$.

1. Démontrer que $f(x) := \frac{1}{2}\|Ax - y\|^2 = \frac{1}{2}\langle Qx|x\rangle + \langle b|x\rangle + c$, où $Q \in \mathcal{M}_N(\mathbb{R})$, $b \in \mathbb{R}^N$ et $c \in \mathbb{R}$ sont à déterminer.
2. En déduire l'expression de $\nabla f(x)$ et $\mathrm{D}^2 f(x)$ en fonction de A et y.
3. Montrer que Q est symétrique, $Q \succeq 0$, et que $x^* \in \arg\min f$ si et seulement si $A^T A x^* = A^T y$, où A^T est la transposée de A.
4. On suppose que A est injective (i.e. son noyau est réduit à 0). Démontrer que Q est définie positive, puis que le problème d'optimisation admet une unique solution x^*.

II. Régression logistique

On étudie les aspects numériques de la régression logistique, un outil très utilisé pour des tâches de classification de données. On considère un jeu de données constitué de points $x_a \in \mathbb{R}^d$, $a \in A = \{1, \ldots, n\}$, que l'on suppose regroupés en deux catégories disjointes $A = A_0 \sqcup A_1$. On posera $y_a = 0$ si $a \in A_0$ et $y_a = 1$ si $a \in A_1$. L'objectif de la classification supervisée est de construire une fonction $u : \mathbb{R}^d \to \mathbb{R}$ permettant d'estimer la catégorie d'un point: on souhaite que $u(x_a) \simeq y_a$ pour tout $a \in A$. Dans la régression logistique, on cherchera une fonction de la forme[1] $u_w(x) = \sigma(\langle w|x\rangle)$ où $w \in \mathbb{R}^d$ est un vecteur et où $\sigma : t \in \mathbb{R} \mapsto e^t/(1 + e^t)$ est la sigmoïde.

[1] Les fonctions de la forme $x \mapsto u_w(x)$ sont bien adaptées pour approcher nos données $(x_a, y_a)_{a\in A}$ où $y_a \in \{0, 1\}$, car elles vérifient automatiquement $u_w(x) \in]0, 1[$ (au contraire des polynômes, qui, lorsqu'ils sont non constant sont aussi non-bornés!).

Problème d'optimisation Informellement, le problème de régression logistique consiste à trouver $w \in \mathbb{R}^d$ tel que pour tout $a \in A$, $\sigma(\langle w|x_a\rangle) \simeq y_a$. On le formalise par le problème d'optimisation suivant, d'inconnue $w \in \mathbb{R}^d$:

$$\min_{w\in\mathbb{R}^d} F(w)$$
$$\text{où } F(w) = \sum_{a\in A} f_a(w) + \frac{\gamma}{2}\|w\|^2 \text{ et } f_a(w) = \begin{cases} -\log(1-\sigma(\langle w|x_a\rangle) \text{ si } a \in A_0 \\ -\log(\sigma(\langle w|x_a\rangle)) \text{ si } a \in A_1 \end{cases} \tag{2}$$

En pratique, une fois trouvé le minimiseur w^* de (2), on a fini la phase d'apprentissage, et on pourra utiliser $u^* := u_{w^*}$ pour estimer la catégorie de points ne faisant pas partie du jeu de données. La figure 1 représente un exemple de données $(x_a)_{a\in A}$ et la fonction u^* reconstruite.

Q0. Montrer que $\sigma(t) \in \,]0,1[$ pour tout $t \in \mathbb{R}$, $\lim_{t\to-\infty}\sigma(t) = 0$ et $\lim_{t\to\infty}\sigma(t) = 1$, et interpréter le problème d'optimisation (2) lorsque $\gamma = 0$.

1 Étude théorique du problème régularisé ($\gamma > 0$)

1.1 Existence et unicité

Q1.[Existence de solutions] Montrer que dans (2), $f_a \geq 0$ pour tout $a \in A$. En déduire que $F(w) \geq \frac{\gamma}{2}\|w\|^2$, puis que F admet un minimiseur sur $\mathbb{R}^d$.

Q2.[Convexité et unicité]
1. Soit $g : \mathbb{R} \to \mathbb{R}$ convexe et $x \in \mathbb{R}^d$. Montrer que $f : w \in \mathbb{R}^d \mapsto g(\langle w|x\rangle)$ est convexe.
2. Soit $g : t \mapsto -\log(1-\sigma(t))$. Montrer que $\sigma'(t) = \sigma(t)(1-\sigma(t))$, en déduire que $g''(t) \geq 0$, que g est convexe et que pour tout $a \in A_0$, f_a est convexe.
   (On montrerait de même que f_a est convexe $\forall a \in A_1$)
3. En déduire que F est γ-fortement convexe *(et que : $\forall w \in \mathbb{R}^d$, $\mathrm{D}^2F(w) \succeq \gamma\mathrm{Id}$.)*
4. Démontrer que le minimiseur de F sur $\mathbb{R}^d$ est unique.

On notera désormais w^* l'unique minimiseur de F.

1.2 Calcul du gradient et de la hessienne

Q3.[Calcul du gradient] Montrer que

$$\begin{aligned}\nabla f_a(w) &= (\sigma(\langle w|x_a\rangle) - y_a)x_a \\ \nabla F(w) &= \sum_{a\in A}(\sigma(\langle w|x_a\rangle) - y_a)x_a + \gamma w,\end{aligned} \tag{3}$$

où l'on rappelle que $y_a = 0$ si $a \in A_0$, $y_a = 1$ si $a \in A_1$ et $\sigma'(t) = \sigma(t)(1-\sigma(t))$.

Q4.[Calcul de la Hessienne]
1. Soit $x \in \mathbb{R}^d$ un vecteur colonne. Montrer que $A = xx^T$ est une matrice carrée symétrique positive dont les entrées sont $(xx^T)_{ij} = x_ix_j$, et que $\langle Av|v\rangle = \langle x|v\rangle^2$.
2. Soit $(x,y) \in \mathbb{R}^d \times \mathbb{R}$, et $G(w) := (\sigma(\langle w|x\rangle) - y)x$. Montrer que

$$\forall i,j \in \{1,\ldots,d\}, \quad \frac{\partial G_i}{\partial w_j}(w) = \sigma(\langle w|x\rangle)(1-\sigma(\langle w|x\rangle))x_ix_j,$$

où l'on a noté $G_i(w)$ la $i$ème coordonnée de $G(w)$. En déduire que

$$\begin{aligned} D^2 f_a(w) &= \sigma(\langle w|x_a\rangle)(1-\sigma(\langle w|x_a\rangle))x_a x_a^T, \\ D^2 F(w) &= \sum_{a\in A} \sigma(\langle w|x_a\rangle)(1-\sigma(\langle w|x_a\rangle))x_a x_a^T + \gamma I_d. \end{aligned} \tag{4}$$

Q5.[Convergence de la descente de gradient]

1. Montrer que, pour $w^{(0)} \in \mathbb{R}^d$, l'ensemble $S = \{w \in \mathbb{R}^d \mid F(w) \le F(w^{(0)})\}$ est compact.
2. En utilisant(4), montrer que pour tout $v \in \mathbb{R}^d$,

$$\langle D^2 f_a(w)v|v\rangle \le \|x_a\|^2 \|v\|^2 .$$

   En déduire que pour tout $w \in \mathbb{R}^d$, $\langle D^2 F(w)v|v\rangle \le M \|v\|^2$, où

$$M = \left(\sum_{a\in A} \|x_a\|^2\right) + \gamma.$$

3. Déduire de ce qui précède que $\forall w \in \mathbb{R}^d$, $\gamma \mathrm{Id} \le D^2 F(w) \le M\mathrm{Id}$.
4. En conclure que les itérées de l'algorithme de descente de gradient à pas optimal (resp. avec backtracking d'Armijo) convergent vers w^*.

2 Étude théorique du problème non régularisé ($\gamma = 0$)

On suppose dorénavant que $\gamma = 0$. Il n'est alors plus évident que F est strictement convexe ni même qu'elle tend vers $+\infty$ lorsque $\|w\| \to +\infty$. On doit donc rajouter une hypothèse sur les points $(x_a)_{a\in A}$. Dans la suite, on l'hypothèse que tout point de $\mathbb{R}^d$ peut être écrit comme une combinaison linéaire à coefficients positifs des points $(x_a)_{a\in A_0}$:

$$\forall w \in \mathbb{R}^d, \exists (\lambda_a)_{a\in A_0} \text{ tq } \lambda_a \ge 0 \text{ et } w = \sum_{a\in A_0} \lambda_a x_a \tag{5}$$

On pose $C = F(w^{(0)})$ où $w^{(0)} \in \mathbb{R}^d$ et $S = \{w \in \mathbb{R}^d \mid F(w) \le C\}$. Pour montrer que le minimum dans (2) est atteint, il suffit donc de démontrer que S est compact.

Q7.[Existence]

1. En utilisant la positivité des f_a, démontrer que si $w \in S$, alors $f_a(w) \le C$. En déduire que

$$\forall a \in A_0, -\log\left(\frac{1}{1+e^{\langle w|x_a\rangle}}\right) \le C$$

   En utilisant la décroissance de $-\log$, en déduire que $\forall a \in A_0, \langle w|x_a\rangle \le C$.
2. (**) Montrer que l'hypothèse (5) implique que

$$\min_{\|w\|=1} \max_{a\in A_0} \langle w|x_a\rangle > 0. \tag{6}$$

   Indication: Poser $\phi(w) = \max_{a\in A_0}\langle w|x_a\rangle$, qui est continue, et raisonner par l'absurde. Montrer que si (6) est fausse, alors il existe $w^* \in \mathbb{R}^d$ tel que $\|w^*\| = 1$ et $\phi(w^*) \le 0$. En utilisant (5), écrire $w^* = \sum_i \lambda_i x_i$ où $\lambda_i \ge 0$. Démontrer qu'alors $\langle w^*|w^*\rangle \le 0$, conclure.

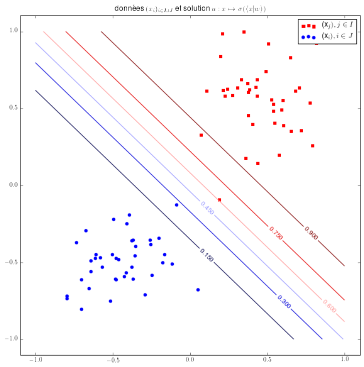

Figure 1: Les points $(x_a)_{a\in A_0}$ sont représentés par des carrés, les points $(x_a)_{a\in A_1}$ par des disques. On trace par dessus des lignes de niveau de la fonction $u_{w^*} : x \mapsto \langle x|w^*\rangle$ où w^* minimise (2).

3. Déduire de (6) qu'il existe $\kappa > 0$ tel que $\forall w \in \mathbb{R}^d, \exists a \in A, \langle w|x_a\rangle \geq \kappa \|w\|$.
4. Conclure que $S \subseteq \{w \in \mathbb{R}^d \mid \|w\| \leq C/\kappa\}$ est compact, puis que F admet un minimiseur.

Q8.[Forte convexité, $\mathrm{D}^2 F > 0$] En utilisant la formule (4) et l'hypothèse (5), démontrer que $\mathrm{D}^2 F(w)$ est symétrique définie positive en tout point $w \in \mathbb{R}^d$.

Q9.[Convergence] En déduire la convergence de la méthode de descente de gradient avec rebroussement d'Armijo.

III. Projection sur un convexe

Le but ici est d'utiliser les théorèmes du cours (notamment le théorème de projection sur un convexe fermé) pour calculer des projections et écrire des conditions d'optimalité pour des problèmes d'optimisation sous contrainte.

3 Projection sur un demi-espace

Q1. Soit $v \in \mathbb{R}^n$ non nul et soit pose $K = \{x \in \mathbb{R}^n \mid \langle v|x\rangle \leq 0\}$. Démontrer que la projection d'un point y sur K est donnée par la formule

$$p_K(y) = \begin{cases} y & \text{si } \langle v|y\rangle \leq 0 \\ y - \langle v|y\rangle \frac{v}{\|v\|^2} & \text{sinon} \end{cases}$$

4 Minimisation avec contrainte de borne inférieure

Dans cette partie, on se donne $K = \{x \in \mathbb{R}^n \mid \forall i \in \{1, \ldots, n\}, x_i \geq f_i\}$ où $f \in \mathbb{R}^n$ est donné. On s'intéresse au problème de minimiser une fonction $J \in \mathcal{C}^1(\mathbb{R}^n)$ sur K. Dans la suite, on note e_i le $i$ème vecteur de la base canonique de $\mathbb{R}^n$.

Q1. Montrer que $p_K(x) = (\max(x_i, f_i))_{1\leq i\leq n}$.

Q2. Soit $J \in \mathcal{C}^1(\mathbb{R}^n)$ et $x^* \in \arg\min_K J$.

1. Si $x_i^* > f_i$, démontrer que $x_\varepsilon := x^* + \varepsilon e_i$ appartient à K pour $\varepsilon \in [-\varepsilon_0, +\infty[$ avec $\varepsilon_0 = x_i^* - f_i$, puis que $J(x_\varepsilon) \geq J(x^*)$. En déduire que $\frac{\partial J}{\partial x_i}(x^*) = 0$.

2. Si $x_i^* = f_i$, démontrer que $\forall \varepsilon \geq 0, x_\varepsilon = x^* + \varepsilon e_i$ appartient à K. En déduire que $\frac{\partial J}{\partial x_i}(x^*) \geq 0$.
Ainsi, $x^* \in \arg\min_K J \Longrightarrow \nabla J(x^*) = \lambda \in \mathbb{R}_+^n$ où $\lambda_i(x_i^* - f_i) = 0$ pour tout $i \in \{1, \ldots, n\}$.

Q3. On suppose que $J \in \mathcal{C}^1(\mathbb{R}^n)$ est convexe. Soit $\bar{x} \in K$ vérifiant $\nabla J(\bar{x}) = \lambda \in \mathbb{R}_+^n$ et $\lambda_i(\bar{x}_i - f_i) = 0$. Démontrer que $\bar{x} \in \arg\min_K J$.

Q4. Dans cette question on considère

$$J : x \in \mathbb{R}^n \mapsto \frac{1}{2}\sum_{i=0}^{n}(x_{i+1} - x_i)^2$$

où l'on pose par convention $x_0 = x_{n+1} = 0$.

1. Calculer $\nabla J(x)$.
2. En utilisant les questions Q2 et Q3, démontrer que

$$\bar{x} \in \arg\min_K J \Longleftrightarrow \bar{x} \in K \text{ et } \begin{cases} \bar{x}_i = \frac{1}{2}(\bar{x}_{i-1} + \bar{x}_{i+1}) & \text{si } \bar{x}_i > f_i \\ \bar{x}_i \geq \frac{1}{2}(\bar{x}_{i-1} + \bar{x}_{i+1}) & \text{si } \bar{x}_i = f_i \end{cases}$$

5 Minimisation sur le simplexe unité

On note $\Delta = \{x \in \mathbb{R}_+^n \mid \sum_{1 \leq i \leq n} x_i = 1\}$, et on appelle cet ensemble Δ le simplexe unité. Cet ensemble apparaît naturellement dans de nombreux problèmes d'optimisation appliqués aux probabilités, en géométrie, ou en finance (allocation de portefeuilles: la contrainte $x_i \geq 0$ dit qu'on investit des sommes positives et la contrainte $\sum_i x_i = 1$ traduit la somme totale que l'on souhaite investir). Dans cet exercice, on cherche à calculer la projection de $y \in \mathbb{R}^n$ sur Δ, i.e.

$$\min_{x \in \Delta} \|x - y\|^2 \tag{7}$$

Pour $\lambda \in \mathbb{R}$, on définit $x^\lambda \in \mathbb{R}^n$ par

$$x_i^\lambda = \max(y_i - \lambda, 0).$$

Notre objectif est de démontrer que la solution de (7) est de la forme x^{λ^*} pour un certain $\lambda^* \in \mathbb{R}$.

Q1. Montrer que le problème (7) admet une unique solution.

Q2. On pose $g(\lambda) = \sum x_i^\lambda$. Montrer que $g(\lambda)$ est continue, que $\lim_{\lambda \to -\infty} g(\lambda) = +\infty$ et $\lim_{\lambda \to +\infty} g(\lambda) = 0$. En déduire qu'il existe λ^* tel que $g(\lambda^*) = 1$, puis que $x^{\lambda^*} \in \Delta$.

Q3. On va montrer que $\bar{x} := x^{\lambda^*}$ est solution de (7). Pour alléger les notations, on pose $\lambda = \lambda^*$.

1. Justifier qu'il suffit de démontrer que

$$\forall z \in \Delta,\ \langle y - \bar{x} | \bar{x} - z \rangle = \sum_{1 \leq i \leq n} (y_i - \bar{x}_i)(\bar{x}_i - z_i) \geq 0.$$

2. Soit $i \in \{1, \ldots, n\}$ tel que $\bar{x}_i > 0$. Démontrer que $\bar{x}_i = y_i - \lambda$ puis que

$$(y_i - \bar{x}_i)(\bar{x}_i - z_i) = \lambda(\bar{x}_i - z_i)$$

3. Soit $i \in \{1, \ldots, n\}$ tel que $\bar{x}_i = 0$. Démontrer alors que $y_i \leq \lambda$ puis que

$$(y_i - \bar{x}_i)(\bar{x}_i - z_i) \geq \lambda(\bar{x}_i - z_i).$$

4. En déduire que $\langle y - \bar{x} | \bar{x} - z \rangle \geq 0$ (en se rappelant $\bar{x}, z \in \Delta$), puis conclure: $\mathrm{p}_\Delta(y) = x^\lambda$.

Q4. Soit $B = \{x \in \mathbb{R}^n \mid \sum_i |x_i| \leq 1\}$ la boule unité pour la norme ℓ^1. On s'intéresse maintenant à la projection d'un point $y \in \mathbb{R}^n$ sur B.

1. Montrer que si $y \in \mathbb{R}^n_+$, alors $p_B(y) \in \mathbb{R}^n_+$.
2. En déduire dans ce cas ($y \in \mathbb{R}^n_+$), que si $y \notin B$, $p_B(y) = p_\Delta(y)$.
3. Montrer que si $y \in \mathbb{R}^n \setminus B$, et si l'on note $y_i = \varepsilon_i \tilde{y}_i$ où $\tilde{y} \in \mathbb{R}^n_+$ et $\varepsilon_i \in \{\pm 1\}$, alors

$$(p_B(y))_i = \varepsilon_i (p_\Delta(\tilde{y}))_i.$$

IV. Application du théorème KKT

Théorème 1. *Soit $J \in \mathcal{C}^1(\mathbb{R}^d)$ convexe, $c_i(x) = \langle a_i | x \rangle - b_i$ pour $1 \leq i \leq k$ et soit*

$$K = \{x \in \mathbb{R}^d \mid \forall i \in \{1, \dots, k\},\ c_i(x) \leq 0.\}$$

Alors,

$$x^* \in \arg\min_K J \iff \exists \lambda \in \mathbb{R}^k, \quad \begin{cases} -\nabla J(x^*) = \sum_{1 \leq i \leq k} \lambda_i \nabla c_i(x^*) & \textit{(équilibre)} \\ \forall 1 \leq i \leq k, c_i(x^*) \leq 0 & \textit{(admissibilité de } x\textit{)} \\ \forall 1 \leq i \leq k, \lambda_i \geq 0 & \textit{(admissibilité de } \lambda\textit{)} \\ \forall 1 \leq i \leq k, \lambda_i c_i(x^*) = 0 & \textit{(complémentarité)} \end{cases} \tag{8}$$

Q1. On considère le problème suivant en dimension $d = 2$, avec $k = 2$ contraintes:

$$\begin{cases} J(x) = (x_1 - 2)^2 + 2(x_2 - 1)^2 \\ c_1(x) = x_1 + 4x_2 - 3 \\ c_2(x) = -x_1 + x_2 \end{cases}$$

1. En utilisant le théorème 1, écrire le système d'équation et d'inéquations devant être vérifié par $x^* \in \mathbb{R}^2$, $\lambda \in \mathbb{R}^2$ pour garantir l'optimalité de $x^* \in \arg\min_K J$.
2. Trouver toutes les solutions de ce système. Pour cela, on distinguera les cas:

$$(\lambda_1 = 0, \lambda_2 = 0),\ (\lambda_1 > 0, \lambda_2 = 0),\ (\lambda_1 = 0, \lambda_2 > 0),\ (\lambda_1 > 0, \lambda_2 > 0).$$

   En déduire la valeur de x^*.

Q2. On considère le problème suivant en dimension $d = 2$, avec $k = 2$ contraintes:

$$\begin{cases} J(x) = (x_1 - 2)^2 + 2(x_2 - 1)^2 + 5x_1 x_2 \\ c_1(x) = -x_1 \\ c_2(x) = x_1 - 2x_2 \end{cases}$$

1. Montrer théoriquement que ce problème a au moins une solution.
2. Écrire le système d'équation et d'inéquations devant être vérifié par $x^* \in \mathbb{R}^2$, $\lambda \in \mathbb{R}^2$ pour garantir l'optimalité de $x^* \in \arg\min_K J$.

3. Trouver toutes les solutions de ce système. Pour cela, on distinguera les cas:

$$(\lambda_1 = 0, \lambda_2 = 0),\ (\lambda_1 > 0, \lambda_2 = 0),\ (\lambda_1 = 0, \lambda_2 > 0),\ (\lambda_1 > 0, \lambda_2 > 0).$$

   En déduire la valeur de x^*.
4. Déterminez la ou les solutions du problème.

Q3. Dans la suite, pour tout vecteur $x \in \mathbb{R}^d$ on notera $x \geq 0$ ssi $\forall i, x_i \geq 0$, et $x \geq y$ si $x - y \geq 0$. On se place dans les conditions du théorème 1 et on note A la matrice dont les lignes sont $a_1, \ldots, a_k$ et b le vecteur colonne $(b_1, \ldots, b_k)$.

1. Montrer que $K = \{x \in \mathbb{R}^d \mid Ax \leq b\}$.
2. Montrer que le théorème KKT peut se reformuler par

$$x^* \in \arg\min_K J \iff \exists \lambda \in \mathbb{R}^k, \quad \begin{cases} -\nabla J(x^*) = A^T \lambda \\ Ax \leq b \\ \lambda \geq 0 \\ \forall 1 \leq i \leq k, \lambda_i(\langle a_i | x^* \rangle - b_i) = 0 \end{cases} \tag{9}$$

Q4. On considère $K = \{x \in \mathbb{R}^d \mid Ax \leq b \text{ et } Ex = f\}$ où $A \in \mathcal{M}_{k,d}(\mathbb{R})$, $E \in \mathcal{M}_{\ell,d}(\mathbb{R})$. Montrer que le théorème KKT peut se reformuler par

$$x^* \in \arg\min_K J \iff \exists (\lambda, \mu) \in \mathbb{R}^k \times \mathbb{R}^\ell, \quad \begin{cases} -\nabla J(x^*) = A^T \lambda + E^T \mu \\ Ax \leq b \text{ et } Ex = f \\ \lambda \geq 0 \\ \forall 1 \leq i \leq k, \lambda_i(\langle a_i | x^* \rangle - b_i) = 0 \end{cases} \tag{10}$$

6 Démonstration du théorème KKT pour une unique contrainte

Soit $c : \mathbb{R}^d \to \mathbb{R}$ une fonction convexe de classe $\mathcal{C}^1$ et $K = \{x \in \mathbb{R}^d \mid c(x) \leq 0\}$. L'objectif de ce problème est de montrer que

$$x^* \in \arg\min_K J \iff \exists \lambda \in \mathbb{R}, \quad \begin{cases} c(x^*) \leq 0 \\ \lambda \geq 0 \\ \lambda c(x^*) = 0 \\ -\nabla J(x^*) = \lambda \nabla c(x^*) \end{cases} \tag{11}$$

Q1. Montrons l'implication $\Longleftarrow$ de (11).

1. Montrer que l'implication est vraie si $\lambda = 0$.
2. On suppose $\lambda > 0$, i.e. $c(x^*) = 0$. Montrer que pour tout $x \in K$,

$$J(x) \geq J(x^*) + \langle \nabla J(x^*) | x - x^* \rangle$$
$$0 \geq c(x) - c(x^*) \geq \langle \nabla c(x^*) | x - x^* \rangle.$$

3. Conclure que $\forall x \in K, J(x) \geq J(x^*)$, soit $x^* \in \arg\min_K J$.

On suppose maintenant que c vérifie l'hypothèse suivante, appelée *condition de qualification*:

$$\forall x \in \mathbb{R}^d, c(x) = 0 \Longrightarrow \nabla c(x) \neq 0, \tag{12}$$

et on considère les deux problèmes d'optimisation, où J est une fonction strictement convexe de classe $\mathcal{C}^1$ vérifiant $\lim_{\|x\| \to +\infty} J(x) = +\infty$:

$$P : \min_{x \in K} J(x) \qquad P_\varepsilon : \min_{x \in \mathbb{R}^d} J_\varepsilon(x) \text{ où } J_\varepsilon(x) = J(x) + \frac{1}{\varepsilon} \max(c(x), 0)^2, \tag{13}$$

Q2. Montrer que la condition (12) est vérifiée pour $c(x) = \|x\|^2 - 1$. Quel est l'ensemble K?
Q3. Justifier l'existence et l'unicité du minimiseur x^* (resp. x^*_ε) pour le problème P (resp. P_ε).
Q4. Montrer que x^*_ε vérifie la condition d'optimalité

$$\nabla J(x^*_\varepsilon) + \lambda_\varepsilon \nabla c(x^*_\varepsilon) = 0, \text{ où } \lambda_\varepsilon = \frac{2}{\varepsilon} \max(c(x_\varepsilon), 0). \tag{14}$$

On admettra dans la suite qu'il existe une suite $\varepsilon_i \to 0$ telle que $\lim_{\varepsilon \to 0} x^*_{\varepsilon_i} = x^*$.
Q5. Dans cette question, on démontre l'implication $\Longrightarrow$ de (11):

1. On suppose $c(x^*) < 0$. Montrer que $\lambda_{\varepsilon_i} = 0$ pour i suffisamment grand, puis en utilisant (14) que $-\nabla J(x^*) = 0 = \lambda \nabla c(x^*)$ où $\lambda = 0$.
2. On suppose maintenant que $c(x^*) = 0$, soit par (12), $-\nabla c(x^*) \neq 0$. Déduire de (14) que si $\varepsilon_i \to 0$, alors $(\lambda_{\varepsilon_i})_i$ est bornée. En déduire l'existence de λ vérifiant (11).