Introduction à l'optimisation convexe non différentiable.

Aude Rondepierre

Table des matières

Introduction

L'objet de ce cours est la résolution de problèmes d'optimisation dont la fonction objectif n'est pas continûment différentiable. Nous nous intéressons ainsi à des problèmes de la forme :

$$\min_{x\in\mathbb{R}^n} f(x),$$

où $f : \mathbb{R}^n \to \mathbb{R}$ est supposée localement Lipschitzienne. Dans ce cours, nous nous intéressons essentiellement à la classe des fonctions convexes. Ces hypothèses requièrent un outil spécial : le sous-différentiel de l'analyse convexe, noté $\partial f(x)$.

Les fonctions convexes apparaissent abondamment dans l'ingénierie et permettent de modéliser de nombreux phénomènes non linéaires (équations de la physique, traitement du signal, théorie des jeux et de l'économie, statistiques...). Elles ont des propriétés remarquables qui permettent d'analyser plus facilement leurs propriétés et aussi de les minimiser efficacement. De nombreux problèmes non convexes irrésolvables peuvent être approchés par des problèmes convexes qui eux sont presque systématiquement minimisés en des temps rapides.

Pour conclure ce cours, nous indiquons quelques références utiles pour aller plus loin. Les livres [8, 24] sont de très bonnes références pour découvrir les nombreux détails de l'analyse convexe en dimension finie. La référence [1] propose est une très bonne introduction à l'analyse convexe dans des espaces de Hilbert. Enfin, indiquons une référence dans les espaces de Banach [27]. Notons que cette référence apporte des informations intéressante même pour des questions d'analyse numérique en dimension finie. Au niveau algorithmique, ce cours est essentiellement inspiré des références [21, 20].

Remerciements : ces notes de cours sont très largement inspirées du cours de Pierre Weiss. Qu'il en soit ici remercié !

Chapitre 1

Eléments d'analyse convexe

Dans ce chapitre, nous présentons quelques propriétés remarquables des fonctions convexes. Elle permettront de construire des algorithmes de minimisation dans la suite du cours.

Dans toutes ces notes, on se place sur l'espace vectoriel $\mathbb{R}^n$, $n \in \mathbb{N}$. On le munit d'un produit scalaire $\langle \cdot, \cdot \rangle$. La norme associée au produit scalaire est notée $\| \cdot \|_2$. Elle est définie pour tout $x \in \mathbb{R}^n$ par $\|x\|_2 = \sqrt{\langle x, x \rangle}$.

1.1 Définitions et propriétés géométriques élémentaires

Une différence importante par rapport aux chapitres précédents est qu'on autorise ici les fonctions à valoir $+\infty$ (mais pas $-\infty$). Ainsi les fonctions considérées dans ce chapitres seront de la forme :

$$f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}.$$

Autoriser les fonctions à valoir $+\infty$ présente l'intérêt de pouvoir représenter les problèmes contraints sous une forme non contrainte. On a en effet :

$$\inf_{x \in \mathbb{R}^n} f(x) = \inf_{x \in dom(f)} f(x)$$

où $dom(f)$ est défini de la façon suivante :

Définition 1.1 (Domaine d'une fonction) *Le domaine d'une fonction f est noté $dom(f)$. Il est défini par*

$$dom(f) = \{x \in \mathbb{R}^n, f(x) < +\infty\}.$$

Dans toute la suite de ce cours, on supposera (sans le préciser) que nos fonctions n'ont pas un domaine vide : $dom(f) \neq \emptyset$.

Définition 1.2 (Ensemble convexe) *Soit $X \subseteq \mathbb{R}^n$ un ensemble. Ce dernier est dit convexe si :*

$$\forall (x_1, x_2) \in X \times X,\ \forall \alpha \in [0,1],\ \alpha x_1 + (1-\alpha) x_2 \in X.$$

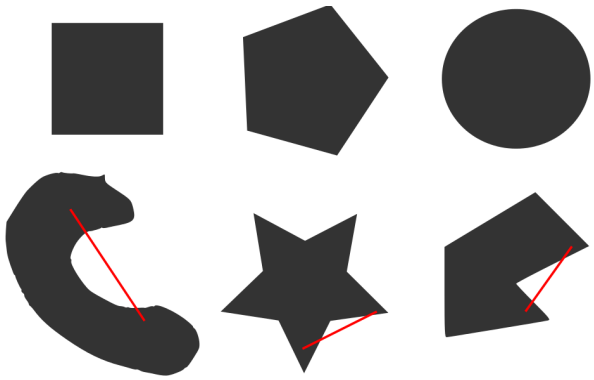


FIGURE 1.1 – En haut : quelques exemples d'ensembles convexes en 2 dimensions. En bas : quelques exemples d'ensembles non convexes (notez qu'il existe des segments dont les extrémités appartiennent à l'ensemble, qui ne sont pas entièrement contenus dans les ensembles).

La définition de la convexité reste identique pour les fonctions à valeur dans $\mathbb{R} \cup \{+\infty\}$.

Définition 1.3 (Fonction convexe) *Soit* $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. f *est dite convexe si :*

$$\forall (x, y) \in dom(f) \times dom(f),\ \forall \lambda \in [0, 1],\ f((1-\lambda)x + \lambda y) \leq (1-\lambda)f(x) + \lambda f(y).$$

Cette définition est illustrée sur la figure 1.1. Elle implique que le domaine d'une fonction convexe est convexe (exercice).

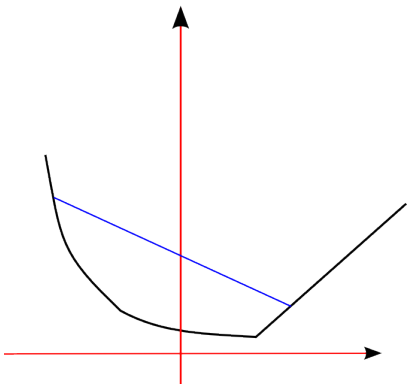


FIGURE 1.2 – Un exemple de fonction convexe. Un segment tracé entre deux points de son graphe reste au dessus du graphe. Notez qu'une fonction convexe peut ne pas être dérivable. On peut cependant montrer qu'elle est dérivable presque partout.

Définition 1.4 (Combinaison convexe) *Soient $x_1, .., x_m$, m éléments de $\mathbb{R}^n$. On dit que x est combinaison convexe de ces points s'il existe $(\alpha_1, ..., \alpha_m)$ tels que :*

- $\forall i \in \{1, ..., m\},\ \alpha_i \in \mathbb{R}_+$,
- $\sum_{i=1}^m \alpha_i = 1$,
- $x = \sum_{i=1}^m \alpha_i x_i$.

Définition 1.5 (Enveloppe convexe) *Soit $X \subseteq \mathbb{R}^n$ un ensemble. On appelle enveloppe convexe de X et on note $conv(X)$ l'ensemble convexe le plus petit contenant X. En dimension finie, c'est aussi l'ensemble des combinaisons convexes d'éléments de X :*

$$conv(X) = \{x \in \mathbb{R}^n \mid x = \sum_{i=1}^{p} \alpha_i x_i \text{ où } x_i \in X,\ p \in \mathbb{N} \text{ et } \sum_{i=1}^{p} \alpha_i = 1,\ \alpha_i \geq 0\}.$$

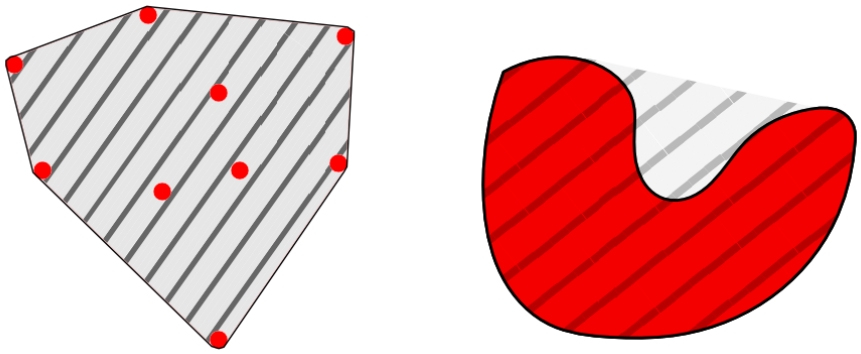


FIGURE 1.3 – Exemples d'enveloppes convexes. A gauche : enveloppe convexe d'un ensemble discret. A droite : enveloppe convexe d'un ensemble continu.

La définition de la convexité permet d'obtenir un lemme souvent utile (notamment en probabilités) :

Lemme 1.1 (Inégalité de Jensen) *Soit $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ une fonction Soient $x_1, x_2, ..., x_m$, m points appartenant à $dom(f)$ et $\alpha_1, ..., \alpha_m$ des coefficients réels positifs tels que $\sum_{i=1}^m \alpha_i = 1$. Alors*

$$f\left(\sum_{i=1}^{m} \alpha_i x_i\right) \leq \sum_{i=1}^{m} \alpha_i f(x_i)$$

Preuve. Par récurrence, on vérifie d'abord que pour $m = 2$, l'inégalité n'est rien d'autre que la définition de la convexité. Puis on suppose le résultat vrai au rang k et on le montre au rang $k + 1$ en réutilisant l'inégalité de convexité. □

Corollaire 1.1 *Soit $x \in \mathbb{R}^n$ une combinaison convexe de $x_1, ..., x_m$ alors $f(x) \leq \max_{i\in\{1,...,m\}} f(x_k)$*

Preuve.

$$f(x) \leq \sum_{i=1}^{m} \alpha_k f(x_k) \leq \max_{i\in\{1,...,m\}} f(x_k).$$

□

Théorème 1.1 *f est convexe si et seulement si son épigraphe*

$$epi(f) = \{(x,t) \in dom(f) \times \mathbb{R},\ t \geq f(x)\}$$

est convexe.

Preuve. Si $(x_1, t_1) \in epi(f)$ et $(x_2, t_2) \in epi(f)$ alors pour tout $\alpha \in [0,1]$ on a

$$\alpha t_1 + (1-\alpha)t_2 \geq \alpha f(x_1) + (1-\alpha) f(x_2) \geq f(\alpha x_1 + (1-\alpha)x_2).$$

Ainsi, $(\alpha x_1 + (1-\alpha)x_2, \alpha t_1 + (1-\alpha)t_2) \in epi(f)$.

Réciproquement, si $epi(f)$ est convexe, alors pour x_1, x_2 dans $dom(f)$, on a

$$(x_1, f(x_1)) \in epi(f),\ (x_2, f(x_2)) \in epi(f).$$

Ainsi, $(\alpha x_1 + (1-\alpha)x_2, \alpha t_1 + (1-\alpha)t_2) \in epi(f)$, soit encore

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2).$$

□

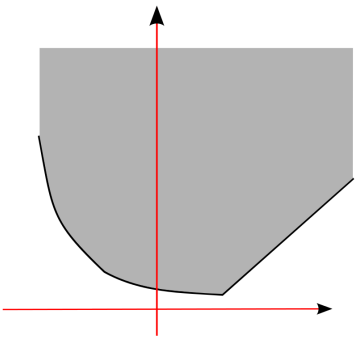


FIGURE 1.4 – L'épigraphe de la fonction est la zone grisée au-dessus du graphe de la fonction (en noir).

Théorème 1.2 *Si f est convexe alors ses ensembles de niveaux*

$$\mathcal{L}_f(\beta) = \{x \in dom(f), f(x) \leq \beta\}$$

sont soit vides, soit convexes.

Preuve. si x_1 et x_2 appartiennent à $\mathcal{L}_f(\beta)$, alors $\forall \alpha \in [0,1]$,

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2) \leq \alpha\beta + (1-\alpha)\beta = \beta.$$

□

Remarque 1.1 *La réciproque est fausse ! Une fonction dont les ensembles de niveaux sont convexes n'est pas convexe en général (exemple en 1D : $f(x) = \sqrt{|x|}$). Une telle fonction est appelée quasi-convexe. Un exemple de fonction quasi-convexe (non convexe) est donné sur la figure 1.1.*

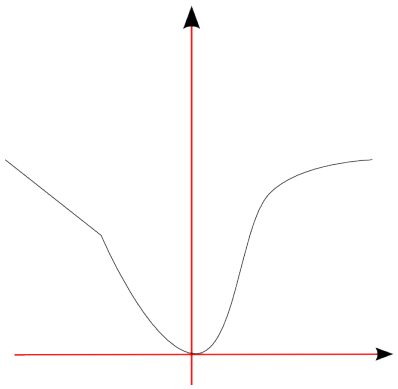


FIGURE 1.5 – Un exemple de fonction quasi-convexe. Ses lignes de niveaux sont des segments (donc des ensembles convexes), mais la fonction n'est pas convexe.

Les fonctions convexes différentiables ont plusieurs caractérisations :

Théorème 1.3 (Caractérisation différentielle de la convexité)
Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable. Les propositions suivantes sont équivalentes :

(a) f est convexe.

(b) Ses hyperplans tangeants sont des minorants (voir figure 1.1).

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

(c) Le gradient de f est un opérateur monotone :

$$\forall (x,y) \in \mathbb{R}^n \times \mathbb{R}^n, \ \langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0.$$

Note : un opérateur $F : \mathbb{R}^n \to \mathbb{R}^n$ est dit monotone si

$$\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, \ \langle F(y) - F(x), y - x \rangle \geq 0.$$

Cette notion généralise la notion de gradient des fonctions convexes. Elle apparaît abondamment dans les EDP.

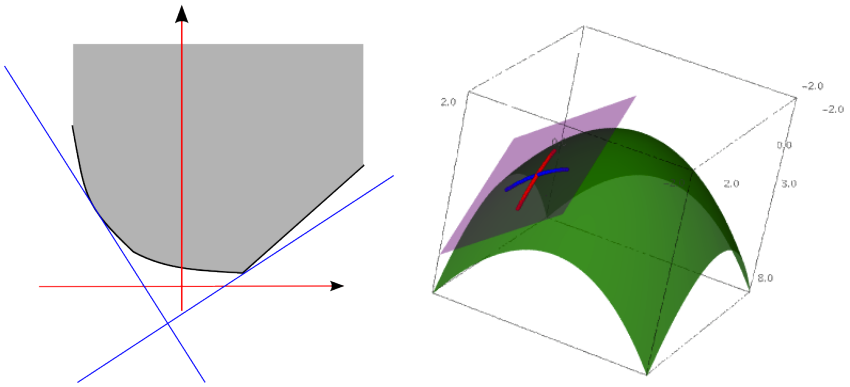


FIGURE 1.6 – Hyperplans tangents d'une fonction convexe (1D) et d'une fonction concave (2D). Notez que l'hyperplan tangent est un minorant pour la fonction convexe et un majorant pour la fonction concave.

Preuve. Nous nous contentons de la preuve $(a) \Leftrightarrow (b)$ et laissons l'équivalence $(b) \Leftrightarrow (c)$ en exercice. On montre d'abord $(a) \Rightarrow (b)$.

Soit $\alpha \in [0, 1]$ et $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$. On a

$$\lim_{\alpha \to 0^+} \frac{f(x + \alpha(y - x)) - f(x)}{\alpha} = \langle \nabla f(x), y - x \rangle$$

par définition de la dérivée directionnelle. Par convexité de f, on a de plus

$$f(x + \alpha(y - x)) = f(\alpha y + (1 - \alpha)x) \leq \alpha f(y) + (1 - \alpha) f(x).$$

Donc

$$\frac{f(x + \alpha(y - x)) - f(x)}{\alpha} \leq f(y) - f(x).$$

En passant à la limite quand $\alpha \to 0^+$, on obtient l'inégalité annoncée.

Montrons maintenant $(b) \Rightarrow (a)$. Soient $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$ et $z = \alpha x + (1 - \alpha) y$. On peut écrire :

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

Soit encore :

$$f(x) \geq f(\alpha x + (1 - \alpha) y) + \langle \nabla f(z), (1 - \alpha)(x - y) \rangle$$
$$f(y) \geq f(\alpha x + (1 - \alpha) y) + \langle \nabla f(z), \alpha(y - x) \rangle$$

En multipliant la première inégalité par α, la seconde par $(1-\alpha)$ puis en additionnant les deux, on obtient l'inégalité de convexité. □

Les fonctions C^2 convexes peuvent être caractérisées par leur hessienne.

Proposition 1.1 (Caractérisation de second ordre de la convexité) *Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction C^2. Elle est convexe si et seulement si $\forall x \in \mathbb{R}^n$, $H[f](x) \succeq 0$ (i.e. la hessienne de f est semi-définie positive en tous points).*

De façon générale, une fonction convexe peut être très compliquée sur le bord de son domaine. Par exemple, la fonction 2D

$$f(x,y) = \begin{cases} 0 & \text{si } x^2+y^2 < 1 \\ \phi(x,y) & \text{si } x^2+y^2 = 1 \text{ avec } \phi(x,y) \geq 0 \end{cases} \tag{1.1}$$

est convexe. Cependant, son comportement sur le bord peut être arbitrairement complexe. Minimiser de telles fonctions est en général impossible. Cette remarque nous mène à nous restreindre à la classe de fonctions semi-continues inférieurement :

Définition 1.6 *Une fonction convexe est dite* fermée *ou* semi-continue inférieurement *(s.c.i.) si son épigraphe est une ensemble fermé.*

Exemple 1.1.1 *La fonction définie par*

$$f(x) = \begin{cases} 0 & \textit{si } x \in [0,1[\\ a \geq 0 & \textit{si } x = 1 \end{cases} \tag{1.2}$$

et illustrée sur la figure 1.1.1 n'est fermée que si $a = 0$.

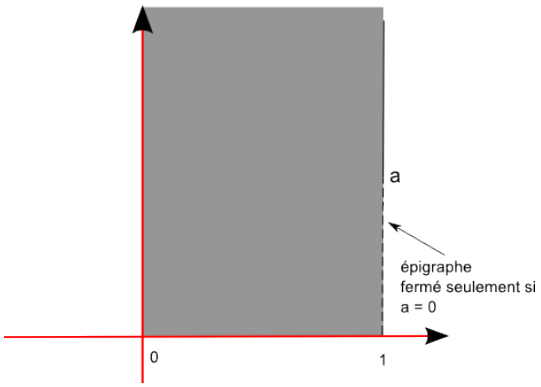


FIGURE 1.7 – Un exemple de fonction convexe dont l'épigraphe est ouvert (voir équation (1.2). Son épigraphe est fermé seulement si $a = 0$.

Remarque 1.2 *On verra plus tard que les fonctions convexes sont continues sur l'intérieur de leur domaine. Toutes les fonctions convexes dont le domaine est $\mathbb{R}^n$ tout entier sont donc continues.*

Théorème 1.4 *Les ensembles de niveau des fonctions convexes s.c.i. sont fermés.*

Exemple 1.1.2 *Donnons quelques exemples de fonctions convexes fermées.*

1. *Les fonctions $f : \mathbb{R}^n \to \mathbb{R}$ convexes telles que $dom(f) = \mathbb{R}^n$ sont fermées. Voir théorème 1.7.*
2. *Les fonctions linéaires sont convexes et fermées.*
3. *Les fonctions convexes, différentiables sur $\mathbb{R}^n$ sont convexes et fermées.*
4. *La fonction $f(x) = \|x\|$ où $\|\cdot\|$ est une norme quelconque est convexe et fermée. En effet, $f(\alpha x_1 + (1-\alpha)x_2) = \|\alpha x_1 + (1-\alpha)x_2\| \leq \alpha\|x_1\| + (1-\alpha)\|x_2\|$.*
5. *La fonction (1.1) n'est convexe et fermée que si $\phi(x, y) = 0,\ \forall (x, y) \in \mathbb{R}^2$.*

1.2 Opérations sur les fonctions convexes

Théorème 1.5 *Soient f_1 et f_2 deux fonctions convexes s.c.i. et $\beta > 0$. Les fonctions suivantes sont convexes s.c.i. :*

1. *$f(x) = \beta f_1(x)$.*
2. *$f(x) = (f_1 + f_2)(x)$ et $dom(f) = dom(f_1) \cap dom(f_2)$.*
3. *$f(x) = \max\{f_1(x), f_2(x)\}$ et $dom(f) = dom(f_1) \cap dom(f_2)$.*

Preuve. Exercice. □

Théorème 1.6 *Soit $\phi(y),\ y \in \mathbb{R}^m$ une fonction convexe s.c.i. Soit $A : \mathbb{R}^n \to \mathbb{R}^m$ un opérateur linéaire et $b \in \mathbb{R}^n$. Alors la fonction $f(x) = \phi(Ax + b)$ est convexe s.c.i. et $dom(f) = \{x \in \mathbb{R}^n, Ax + b \in dom(\phi)\}$.*

Preuve. Soient $(x_1, x_2) \in dom(f) \times dom(f)$. On note $y_1 = Ax_1 + b$ et $y_2 = Ax_2 + b$. Alors pour tout $\alpha \in [0, 1]$:

$$\begin{aligned} f(\alpha x_1 + (1-\alpha)x_2) &= \phi(\alpha(Ax_1 + b) + (1-\alpha)(Ax_2 + b)) \\ &\leq \alpha\phi(Ax_1 + b) + (1-\alpha)\phi(Ax_2 + b) \\ &\leq \alpha f(x_1) + (1-\alpha)f(x_2). \end{aligned}$$

La fermeture de l'épigraphe est due à la continuité de l'opérateur affine $x \mapsto Ax + b$. □

1.3 Continuité et différentiabilité

Lemme 1.2 *Soit f une fonction convexe et $x_0 \in int(dom(f))$. Alors f est majorée localement en x_0* [1].

1. Note : elle peut exploser au voisinage du bord. Il suffit par exemple de considérer la fonction $x \mapsto 1/x$ sur $]0, 1]$ pour s'en convaincre.

Preuve. Soit $\epsilon > 0$ tel que les points $x_0 \pm \epsilon e_i \in int(dom(f))$. D'après le corollaire 1.1 on a $\forall x \in conv(\{x_0 \pm \epsilon e_i\}) = B_\infty(x_0, \epsilon)$, $f(x) \leq \max_{1 \leq i \leq n} f(x_0 \pm \epsilon e_i) < +\infty$. □

Remarque 1.3 *Dans ces notes de cours, nous énonçons en général les résultats sur $int(dom(f))$. En réalité, la grande majorité des résultats présentés sont valables sur l'intérieur relatif de $dom(f)$, c'est-à-dire l'intérieur par rapport à la topologie induite sur le plus petit sous-espace affine contenant $dom(f)$. Par exemple, l'intérieur du simplexe sur $\mathbb{R}^n$ est vide alors que l'intérieur relatif du simplexe est l'ensemble $\{x \in \mathbb{R}^n, x_i > 0, \forall i \in \{1, \ldots, n\}, \sum_{i=1}^{n} x_i = 1\}$.*

Le lemme 1.2 implique la continuité d'une fonction convexe sur l'intérieur de son domaine.

Théorème 1.7 *Soit f une fonction convexe et $x_0 \in int(dom(f))$. Alors f est continue et localement Lipschitz en x_0.*

Preuve. D'après le lemme (1.2), il existe $M < +\infty$ et $\epsilon > 0$ tels que $f(x) \leq M, \forall x \in B_2(x_0, \epsilon)$. Soit $y \in B_2(x_0, \epsilon)$ et $z \in \partial B_2(x_0, \epsilon)$ un point de la frontière tel que $z = x_0 + \frac{1}{\alpha}(y - x_0)$ avec $\alpha = \frac{1}{\epsilon}\|y - x_0\|_2$. Par construction $\alpha \leq 1$ et $y = \alpha z + (1-\alpha)x_0$. Par convexité de f on a donc :

$$\begin{aligned} f(y) &\leq (1-\alpha)f(x_0) + \alpha f(z) \\ &\leq f(x_0) + \alpha(M - f(x_0)) \\ &= f(x_0) + \frac{M - f(x_0)}{\epsilon}\|y_\alpha - x_0\|_2. \end{aligned}$$

Pour obtenir l'inégalité inverse, on considère un point $y \in B_2(x_0, \epsilon)$ et on pose $u = x_0 + \frac{1}{\alpha}(x_0 - y)$. On a $\|u - x_0\|_2 = \epsilon$ et $y = x_0 + \alpha(x_0 - u)$. D'après le théorème **??** on a donc :

$$\begin{aligned} f(y) &\geq f(x_0) + \alpha(f(x_0) - f(u)) \\ &\geq f(x_0) - \alpha(M - f(x_0)) \\ &= f(x_0) - \frac{M - f(x_0)}{\epsilon}\|y - x_0\|_2. \end{aligned}$$

On a donc $\forall y \in B_2(x_0, \epsilon)$:

$$|f(y) - f(x_0)| \leq \frac{M - f(x_0)}{\epsilon}\|y - x_0\|_2.$$

□

Définition 1.7 (Dérivée directionnelle) *Soit $x \in dom(f)$. On appelle dérivée directionnelle au point x dans la direction p la limite suivante :*

$$f'(x,p) = \lim_{\alpha \to 0^+} \frac{1}{\alpha}(f(x+\alpha p) - f(x)).$$

Si cette limite existe, on dit que $f'(x,p)$ est la dérivée directionnelle de f en x.

Théorème 1.8 *Une fonction convexe est différentiable dans toutes les directions sur tous les points de l'intérieur de son domaine.*

Preuve. Soit $x \in int(dom(f))$. On considère la fonction

$$\phi(\alpha) = \frac{1}{\alpha}(f(x+\alpha p) - f(x)), \alpha > 0.$$

Soit $\beta \in]0,1]$ et $\alpha \in]0,\epsilon]$ avec ϵ suffisamment petit pour que $x + \epsilon p \in dom(f)$. Alors :

$$f(x+\alpha\beta p) = f((1-\beta)x + \beta(x+\alpha p)) \leq (1-\beta)f(x) + \beta f(x+\alpha p).$$

Ainsi

$$\phi(\alpha\beta) = \frac{1}{\alpha\beta}(f(x+\alpha\beta p) - f(x_0)) \leq \frac{1}{\alpha}(f(x+\alpha p) - f(x)) = \phi(\alpha).$$

La fonction ϕ est donc décroissante au voisinage de 0^+. Pour $\gamma > 0$ tel que $x - \gamma p \in dom(f)$ on a d'après le théorème **??** :

$$\phi(\alpha) \geq \frac{1}{\gamma}(f(x) - f(x-\gamma p)).$$

La limite quand $\alpha \to 0^+$ existe donc. □

Rappelons que la dérivabilité selon toute direction en x n'implique pas nécessairement la différentiabilité de f en x. Le contre-exemple typique est la fonction $x \to |x|$. Cette fonction est bien dérivable à droite et à gauche de 0, mais elle n'est pas dérivable en 0.

Lemme 1.3 *Soit f une fonction convexe et $x \in int(dom(f))$. Alors $f'(x,p)$ est une fonction convexe en p, homogène de degré 1. Pour tout $y \in dom(f)$ on a*

$$f(y) \geq f(x) + f'(x, y-x). \tag{1.3}$$

Preuve. En exercice. □

1.4 Sous-différentiel

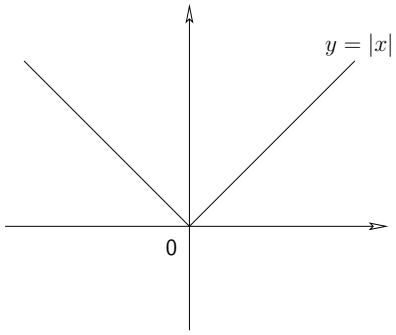
Définition 1.8 (Sous-gradient et sous-différentiel) *Soit f une fonction convexe. Un vecteur $\eta \in \mathbb{R}^n$ est appelé* sous-gradient *de f au point $x_0 \in dom(f)$ si*

$$\forall x \in dom(f),\ f(x) \geq f(x_0) + \langle \eta, x - x_0 \rangle. \tag{1.4}$$

L'ensemble de tous les sous-gradients en x_0 est appelé sous-différentiel *de f. Il est noté $\partial f(x_0)$.*

L'interprétation géométrique du sous-différentiel est la suivante. Il est formé par toutes les directions des hyperplans qui passent par le point $(x, f(x))$ et restent "sous" le graphe de la fonction f. Ces hyperplans sont appelés *hyperplans support* ou *hyperplans d'appui* au graphe de f en x.

Exemple 1.4.1 *Soit $f : x \in \mathbb{R} \mapsto |x|$. Calculons les sous-gradients de f en tout point x de $\mathbb{R}$.*



Dans $\mathbb{R}^2$, les hyperplans d'appui sont des droites et les sous-gradients associés leurs coefficients directeurs.

$$\partial f(x) = \begin{cases} \{-1\} & si \quad x < 0 \\ \{+1\} & si \quad x > 0 \\ [-1, 1] & si \quad x = 0 \end{cases}$$

Lemme 1.4 *Le sous-différentiel $\partial f(x_0) = \{\eta \in \mathbb{R}^n,\ \forall x \in \mathbb{R}^n,\ f(x) \geq f(x_0) + \langle \eta, x - x_0 \rangle\}$ est un ensemble convexe fermé.*

Preuve. Soient g_1 et g_2 des éléments de $\partial f(x_0)$. On a $\forall y \in \mathbb{R}^n$:

$$\begin{aligned} f(y) &\geq f(x) + \langle g_1, y - x \rangle \\ f(y) &\geq f(x) + \langle g_2, y - x \rangle. \end{aligned}$$

Soit $\alpha \in [0, 1]$. En multipliant la première inégalité par α, la deuxième par $(1 - \alpha)$ et en sommant, on voit que $\alpha g_1 + (1 - \alpha) g_2 \in \partial f(x_0)$. □

Théorème 1.9 *Soit f une fonction convexe s.c.i. et $x_0 \in int(dom(f))$. Alors $\partial f(x_0)$ est une ensemble non vide, convexe, borné.*

Preuve. Notons d'abord que le point $(f(x_0), x_0)$ appartient à la frontière de $epi(f)$. D'après le théorème de séparation de Hahn-Banach, il existe donc un hyperplan d'appui à $epi(f)$ au point $(f(x_0), x_0)$:

$$-\alpha\tau + \langle d, x \rangle \leq -\alpha f(x_0) + \langle d, x_0 \rangle \tag{1.5}$$

pour tout $(\tau, x) \in epi(f)$. On peut choisir

$$\|d\|_2^2 + \alpha^2 = 1. \tag{1.6}$$

Puisque pour tout $\tau \geq f(x_0)$ le point (τ, x_0) appartient à $epi(f)$, on conclut que $\alpha \geq 0$.

D'après le lemme 1.2, il existe $\epsilon > 0$ et $M > 0$ tels que $B_2(x_0, \epsilon) \subseteq dom(f)$ et

$$f(x) - f(x_0) \leq M\|x - x_0\|$$

pour tout $x \in B_2(x_0, \epsilon)$. Ainsi, d'après (1.5), pour tout x dans cette boule on a

$$\langle d, x - x_0\rangle \leq \alpha(f(x) - f(x_0)) \leq \alpha M\|x - x_0\|_2.$$

En choisissant $x = x_0 + \epsilon d$ on obtient $\|d\|_2^2 \leq M\alpha\|d\|_2$. En utilisant la condition de normalisation (1.6), on obtient

$$\alpha \geq \frac{1}{\sqrt{1 + M^2}}.$$

Ainsi, en choisissant $g = d/\alpha$ on obtient pour tout $x \in dom(f)$

$$f(x) \geq f(x_0) + \langle g, x - x_0\rangle.$$

Finalement, si $g \in \partial f(x_0)$ et $g \neq 0$, alors en choisissant $x = x_0 + g/\|g\|_2$ on obtient

$$\epsilon\|g\|_2 = \langle g, x - x_0\rangle \leq f(x) - f(x_0) \leq M\|x - x_0\|_2 = M\epsilon.$$

Ainsi $\partial f(x_0)$ est borné. □

Remarque 1.4 *Le sous-différentiel peut ne pas exister sur le bord du domaine. Par exemple la fonction* $f(x) = -\sqrt{x}$ *sur* $\mathbb{R}_+$ *est convexe et fermée, mais le sous-différentiel n'existe pas en* 0 *car* $\lim_{x\to 0^+} f(x) = -\infty$.

Le sous-différentiel est central car il permet notamment de caractériser les minimiseurs d'une fonction. On considère le problème suivant :

$$\text{Trouver } \; x^* \text{ tel que } f(x^*) = \min_{x\in\mathbb{R}^n} f(x)$$

où f est convexe s.c.i.

Théorème 1.10 *x^* est solution du problème ci-dessus si et seulement si* $0 \in \partial f(x^*)$.

Preuve. Si $0 \in \partial f(x^*)$, alors $f(x) \geq f(x^*) + \langle 0, x - x^*\rangle \geq f(x^*),\ \forall x \in dom(f)$. Réciproquement, si $f(x) \geq f(x^*), \forall x \in dom(f)$, alors $0 \in \partial f(^*)$ par définition du sous-différentiel. □

1.5 Règles de calcul sous-différentiel

Lemme 1.5 *Soit f une fonction convexe s.c.i. différentiable sur son domaine. Alors $\forall x \in int(dom(f))$, $\partial f(x) = \{\nabla f(x)\}$.*

Lemme 1.6 *Soit f une fonction convexe s.c.i. avec $dom(f) \subseteq \mathbb{R}^n$. Soit $A : \mathbb{R}^m \to \mathbb{R}^n$ un opérateur linéaire et $b \in \mathbb{R}^n$. La fonction $\phi(x) = f(Ax + b)$ est convexe s.c.i. et $\forall x \in int(dom(\phi)), \partial\phi(x) = A^T \partial f(Ax + b)$.*

Preuve. Exercice. □

Lemme 1.7 *Soit $f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x)$ où f_1 et f_2 sont convexes s.c.i. sur $\mathbb{R}^n$ et $(\alpha_1, \alpha_2) \in \mathbb{R}_+ \times \mathbb{R}_+$. Alors $\partial f(x) = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x) = \{\eta \in \mathbb{R}^n, \exists (x_1, x_2) \in \partial f_1(x_1) \times \partial f_2(x_2), \eta = x_1 + x_2\}$.*

Lemme 1.8 *Soit $f : \mathbb{R}^n \to \mathbb{R}^n \cup \{+\infty\}$ une fonction convexe fermée et $g : \mathbb{R} \cup +\infty \to \mathbb{R} \cup +\infty$ une fonction convexe croissante. Notons $h = g \circ f$. Alors $\forall x \in int(dom(f))$, $\partial h(x) = \{\eta_1 \eta_2, \eta_1 \in \partial g(f(x)), \eta_2 \in \partial f(x)\}$.*

Exemple 1.5.1 *Soit $f(x) = \|x\|_2$, $g(t) = \frac{1}{2}t^2$ et $h(x) = g(f(x)) = \frac{1}{2}\|x\|_2^2$. On peut montrer (exercice) que :*

$$\partial f(x) = \begin{cases} \left\{\frac{x}{\|x\|_2}\right\} & si\ x \neq 0, \\ \{x, \|x\|_2 \leq 1\} & sinon. \end{cases} \tag{1.7}$$

Donc d'après le lemme 1.8 :

$$\partial h(x) = g'(\|x\|_2)\partial f(x) = x, \tag{1.8}$$

et on retrouve le résultat standard.

Lemme 1.9 *Soit $(f_i)_{i=1..m}$ un ensemble de fonctions convexes s.c.i. Alors la fonction*

$$f(x) = \max_{i=1..m} f_i(x)$$

est aussi convexe s.c.i. de domaine $dom(f) = \cap_{i=1}^m dom(f_i)$ et $\forall x \in int(dom(f))$, on a :

$$\partial f(x) = conv\,(\partial f_i(x), i \in I(x))$$

où $I(x) = \{i \in \{1, .., m\}, f_i(x) = f(x)\}$.

Le lemme suivant se révèle souvent utile pour étudier les fonctions définies commes des suprémas [2].

2. En réalité, toutes les fonctions convexes peuvent êtres représentées comme un suprémum en utilisant la transformée de Fenchel. Ce résultat est donc central. Le cours est trop court pour présenter cette théorie.

Lemme 1.10 *Soit $\phi(x,y)$ une fonction telle que $\forall y \in \Delta$, la fonction $x \mapsto \phi(x,y)$ est convexe s.c.i. Alors la fonction*

$$f(x) = \sup_{y\in\Delta} \phi(x,y)$$

est convexe s.c.i.. De plus, $dom(f) = \{x \in \mathbb{R}^n,\ \exists\gamma,\ \phi(x,y) \leq \gamma,\ \forall y \in \Delta\}$ et

$$\partial f(x) \supseteq conv\left(\partial_x\phi(x,y),\ y \in I(x),\ I(x) = \{y, \phi(x,y) = f(x)\}\right).$$

Définition 1.9 *Soit $Q \subseteq \mathbb{R}^n$ un ensemble convexe, fermé. Soit $x_0 \in \partial Q$. On appelle cône normal à Q en x_0 et on note $N_Q(x_0)$ l'ensemble suivant :*

$$N_Q(x_0) = \{\eta \in \mathbb{R}^n, \langle \eta, x - x_0\rangle \leq 0, \forall x \in Q\}.$$

Deux exemples de cône normaux sont donnés sur la figure 1.5.

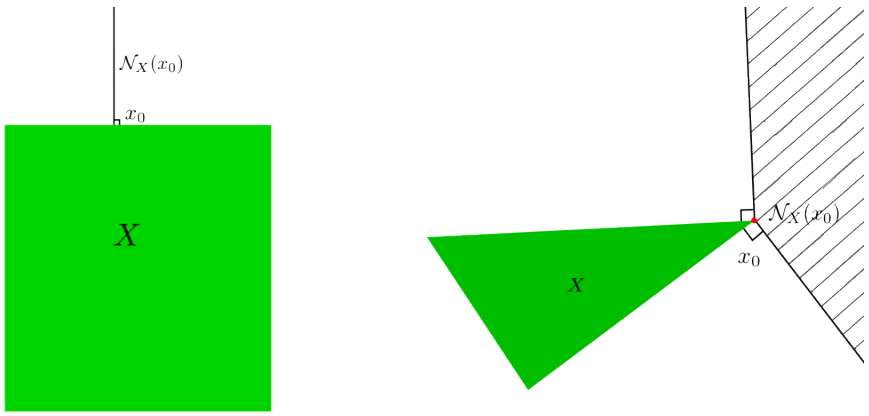


FIGURE 1.8 – Deux exemples de cônes normaux. A gauche : cône normal sur un point régulier de la frontière. A droite : cône normal sur une singularité.

Lemme 1.11 (Sous-différentiel d'une indicatrice) *Soit $X \in \mathbb{R}^n$ un ensemble convexe fermé. On définit l'indicatrice de X comme :*

$$\chi_X(x) = \begin{cases} 0 & si\ \ x \in X \\ +\infty & sinon \end{cases}$$

Le sous-différentiel de χ_X au point $x \in \partial X$ est le cône normal de X en x :

$$\partial\chi_X(x) = N_X(x).$$

Preuve. Exercice. □

Définition 1.10 (Norme duale) *Soit $\|\cdot\|_X$ une norme sur $\mathbb{R}^n$. On appelle norme duale et on note $\|\cdot\|_{X^*}$ la fonction définie par :*

$$\|y\|_{X^*} = \sup_{x\in\mathbb{R}^n, \|x\|_X\leq 1} \langle x, y\rangle. \tag{1.9}$$

On peut montrer que la fonction $\|\cdot\|_{X^*}$ définit bien une norme sur $\mathbb{R}^n$. De plus, par définition, on obtient pour tout $(x, y) \in \mathbb{R}^{n\times n}$:

$$|\langle x, y\rangle| \leq \|x\|_X \|y\|_{X^*}, \tag{1.10}$$

qui généralise les inégalités de Cauchy-Schwartz et de Hölder.

Exemple 1.5.2 *Soit $\|x\|_X = \|x\|_p$ où $\|x\|_p$ est la norme ℓ^p usuelle sur $\mathbb{R}^n$. Alors, on peut montrer que $\|y\|_{X^*} = \|x\|_q$ avec $1/p + 1/q = 1$. En particulier, la norme ℓ^2 est égale à sa norme duale.*

Définition 1.11 (Sous-différentielle d'une norme) *Soit $f(x) = \|x\|_X$ où $\|\cdot\|_X$ est une norme arbitraire sur $\mathbb{R}^n$. Alors :*

$$\partial f(x) = \arg\max_{\|y\|_{X^*}\leq 1} \langle x, y\rangle. \tag{1.11}$$

Preuve. Exercice. □

Pour finir ce paragraphe, montrons que les résultats ci-dessus, et en particulier le théorème (1.9), permettent de retrouver les conditions d'optimalité de Karush-Kuhn-Tucker.

Théorème 1.11 *Soit $(f_i)_{0\leq i\leq m}$ un ensemble de fonctions convexes différentiables telles qu'il existe $\bar{x} \in \mathbb{R}^n$ en lequel les contraintes sont qualifiées.Alors, un point x^* est solution du problème :*

$$\min_{x\in\mathbb{R}^n,\ f_i(x)\leq=0,\ 1\leq i\leq m} f_0(x) \tag{1.12}$$

si et seulement si il existe des nombres $\lambda_i \geq 0$ tels que :

$$\nabla f_0(x^*) + \sum_{i\in I^*} \lambda_i \nabla f_i(x^*) = 0, \tag{1.13}$$

où $I^* = \{i \in \{1, \ldots, m\},\ f_i(x^*) = 0\}$.

Preuve. En exercice. □

1.6 Transformée de Fenchel ou conjuguée convexe

La fonction conjuguée, aussi appelée transformée de Fenchel ou transformée de Legendre-Fenchel est utilisée pour :

- convexifier une fonction (en calculant la bi-conjuguée, i.e. la conjuguée de la conjuguée).
- calculer le sous-différentiel d'une fonction convexe.
- calculer des problèmes dits "duaux", en optimisation. Ces problèmes apportent souvent beaucoup d'information sur les problème "primaux", i.e. ceux que l'on souhaite résoudre.
- passer de la mécanique lagrangienne à la mécanique hamiltonienne,...

Elle a été introduite par Mandelbrojt en 1939 puis précisée et améliorée par Fenchel en 1949. Cette transformée généralise la transformation de Legendre (1787).

Définition 1.12 (Conjuguée convexe) *Soit $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ une fonction. Sa conjuguée est définie par :*

$$f^*(s) = \sup_{x \in \mathbb{R}^n} \langle s, x \rangle - f(x). \tag{1.14}$$

L'application $f \mapsto f^*$ est appelée transformation de Lengendre-Fenchel. La fonction f^* est appelée conjuguée convexe, transformée de Fenchel ou transformée de Legendre-Fenchel de f.

La notion de transformée de Fenchel est illustrée sur la figure 1.6.

La motivation pour introduire cette transformation est la suivante. On peut définir la convexifiée fermée d'une fonction comme l'enveloppe supérieure de toutes les minorantes affines de f. Parmi toutes les minorantes affines $x \mapsto \langle s, x \rangle + \alpha$, on ne peut garder que celle qui est la plus haute, c'est-à-dire qui a le plus grand α. Il faut donc déterminer le plus grand α tel que $\forall x \in \mathbb{R}^n$:

$$\langle s, x \rangle + \alpha \leq f(x).$$

La plus petite valeur de $-\alpha$ est donnée par

$$f^*(s) = \sup_{x \in \mathbb{R}^n} \langle s, x \rangle - f(x).$$

Proposition 1.2 *Pour toute fonction f, f^* est convexe, fermée.*

Preuve. L'intersection d'ensembles fermés est fermé. □

Proposition 1.3 *Pour tout $(x, s) \in \mathbb{R}^n \times \mathbb{R}^n$:*

$$f(x) + f^*(s) \geq \langle x, s \rangle. \tag{1.15}$$

Preuve. Evident d'après la définition. □

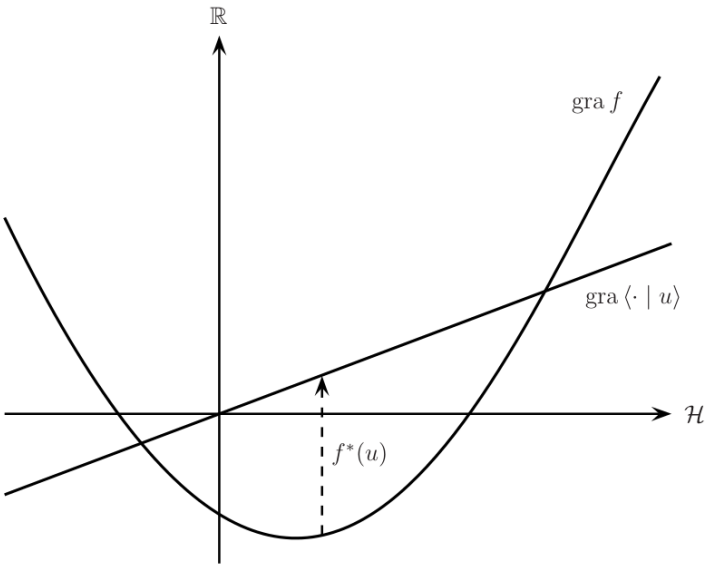


FIGURE 1.9 – $f^*(u)$ est le supremum de la différence signée verticale entre le graphe de f et celui de l'hyperplan défini par la fonction linéaire $\langle \cdot, u \rangle$.

Définition 1.13 (Biconjuguée) *Soit $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ une fonction. Sa biconjuguée est définie par :*

$$f^{**}(x) = \sup_{s \in \mathbb{R}^n} \langle x, s \rangle - f^*(s). \tag{1.16}$$

Théorème 1.12 *La biconjuguée de f est la plus grande fonction convexe fermée inférieure à f. On peut aussi la voir comme la fonction dont l'épigraphe est l'enveloppe convexe fermée de $epi(f)$.*

Preuve. Soit $\Sigma \subset \mathbb{R}^n \times \mathbb{R}$ l'ensemble des paires (s, α) qui définissent une fonction affine $x \mapsto \langle s, x \rangle - \alpha$ majorée par f :

$$\begin{aligned}(s,\alpha) \in \Sigma &\Leftrightarrow f(x) \geq \langle s, x \rangle - \alpha, \ \forall x \in \mathbb{R} \\ &\Leftrightarrow \alpha \geq \sup_{x \in \mathbb{R}^n} \langle s, x \rangle - f(x) \\ &\Leftrightarrow \alpha \geq f^*(x), \ (\text{and } s \in dom(f^*)).\end{aligned}$$

On obtient donc pour $x \in \mathbb{R}^n$,

$$\begin{aligned}\sup_{(s,\alpha)\in\Sigma} \langle s, x \rangle - \alpha &= \sup_{s \in dom(f^*),\ -\alpha \leq -f^*(s)} \langle s, x \rangle - \alpha \\ &= \sup_{s \in dom(f^*)} \langle s, x \rangle - f^*(s) \\ &= f^{**}(x).\end{aligned}$$

D'un point de vue géométrique, les épigraphes des fonctions affines associées à $(s, \alpha) \in \Sigma$ sont les demi-espaces fermés contenant $epi(f)$. L'épigraphe de leur supremum est l'enveloppe convexe fermée de $epi(f)$. □

Théorème 1.13 *La biconjuguée de f satisfait $f^{**} = f$ si et seulement si f est convexe fermée.*

Preuve. C'est une conséquence directe du théorème 1.12. □

Exemple 1.6.1 *Voici quelques exemples de conjuguées convexes. Les preuves sont laissées en exercice.*

- *Soit $p \in]1, +\infty[$ et q le conjugué de p (i.e. tel que $1/p + 1/q = 1$). Alors $(1/p|\cdot|^p)^* = 1/q|\cdot|^q$.*
- *Soit $Q \in \mathbb{R}^{n\times n}$ une matrice SDP et $f(x) = \frac{1}{2}\langle x, Qx\rangle$, alors $f^*(x) = \frac{1}{2}\langle Q^{-1}x, x\rangle$.*
- *Soit L un sous-espace vectoriel de $\mathbb{R}^n$. On considère la fonction :*

$$f(x) = \chi_L(x) = \begin{cases} 0 & si\ x \in L \\ +\infty & sinon \end{cases} \tag{1.17}$$

Sa conjuguée convexe $f^* = \chi_{L^\perp}$.

- *Soit $f(x) = \|x\|$ où $\|\cdot\|$ est une norme quelconque. Alors :*

$$f^*(s) = \chi_B(s) \tag{1.18}$$

où $B = \{s \in \mathbb{R}^n, \|s\|^* \leq 1\}$ et où $\|\cdot\|^*$ est la norme duale à $\|\cdot\|$.

Proposition 1.4 *Soit f une fonction convexe fermée. On a $\forall (x, s) \in \mathbb{R}^n \times \mathbb{R}^n$:*

$$s \in \partial f(x) \Leftrightarrow x \in \partial f^*(s). \tag{1.19}$$

Preuve. On a :

$$\begin{aligned} s \in \partial f(x) &\Leftrightarrow f^*(s) + f(x) = \langle s, x\rangle \\ &\Leftrightarrow f^*(s) + f^{**}(x) = \langle s, x\rangle \\ &\Leftrightarrow x \in \partial f^*(s). \end{aligned}$$

□

1.7 Eléments d'analyse pour l'algorithmie

On pourrait penser qu'une fonction différentiable est plus facile à minimiser qu'une fonction non-différentiable. De même, une fonction strictement convexe pourrait être plus facile à minimiser qu'une fonction simplement convexe (notamment parce qu'elle admet un minimiseur unique). Il n'en est rien. En effet, on peut approcher de façon aussi proche que

l'on souhaite une fonction non différentiable par une fonction C^∞ et une fonction convexe par une fonction strictement convexe. Il faut donc introduire des classes de fonctions plus régulières pour développer des algorithmes efficaces. Les deux classes présentées ci-après sont les plus naturelles : les fonctions différentiables à gradient Lipschitz et les fonctions fortement convexes.

1.7.1 Fonctions à gradient Lipschitz

De façon générale, les fonctions différentiables ne sont pas plus faciles à minimiser que les fonctions non différentiables. On peut en effet approcher n'importe quelle fonction, aussi précisément qu'on le souhaite en norme L^∞, par une fonction C^∞. La différentiabilité n'est donc pas une propriété assez forte pour permettre de faire des estimations d'erreurs sur les méthodes d'optimisation.

C'est pour cette raison qu'on introduit la classe des fonctions différentiables à gradient Lipschitz. Dans tout ce paragraphe, on considère des fonctions $f : \mathbb{R}^n \to \mathbb{R}$ différentiables à gradient Lipschitz :

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|. \tag{1.20}$$

Ces fonctions ont une propriété essentielle pour la preuve de convergence des méthodes de descente :

Lemme 1.12 *Soit $f : \mathbb{R}^n \to \mathbb{R}$ une fonction différentiable à gradient L-Lipschitz continu. Elle satisfait :*

$$f(y) \leq \underbrace{f(x) + \langle \nabla f(x), y - x\rangle}_{\textit{Approximation linéaire}} + \frac{L}{2}\|y - x\|^2$$

Ce résultat indique que pour les fonctions à gradient Lipschitz, l'estimation linéaire d'une fonction ne peut pas être trop mauvaise.

Preuve. On a :

$$\begin{aligned}
f(y) &= f(x) + \int_{t=0}^{1} \langle \nabla f(x + t(y - x)), y - x\rangle dt \\
&= f(x) + \int_{t=0}^{1} \langle \nabla f(x + t(y - x)) - \nabla f(x) + \nabla f(x), y - x\rangle dt \\
&= f(x) + \langle \nabla f(x), y - x\rangle + \int_{t=0}^{1} \langle \nabla f(x + t(y - x)) - \nabla f(x), y - x\rangle dt \\
&\overset{\text{(1.20)+Cauchy-Schwartz}}{\leq} f(x) + \langle \nabla f(x), y - x\rangle + \int_{t=0}^{1} tL\|y - x\|^2 dt \\
&= f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2.
\end{aligned}$$

□

Les fonctions à gradient Lipschitz peuvent être caractérisées par leur hessienne.

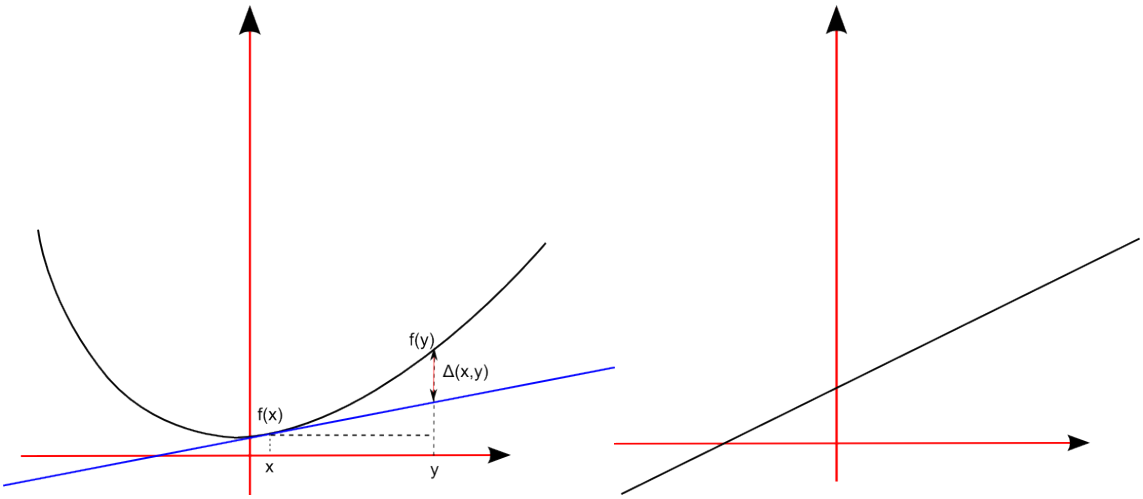


FIGURE 1.10 – Fonction à gradient Lipschitz : le graphe de la courbe ne peut pas s'éloigner trop rapidement de la tangente. Ici $\Delta(x,y) \leq \frac{L}{2}\|y-x\|_2^2$. La fonction affine de droite satisfait $L = 0$ puisque la tangente est identique à la fonction.

Lemme 1.13 *$f : \mathbb{R}^n \to \mathbb{R}$ est une fonction C^2, convexe à gradient L Lipschitz si et seulement si :*

$$\lambda_{min}\left(H[f](x)\right) \geq 0, \forall x \in \mathbb{R}^n. \text{ (convexité)}$$

$$\lambda_{max}\left(H[f](x)\right) \leq L, \forall x \in \mathbb{R}^n. \text{ (gradient Lipschitz).}$$

On finit ce paragraphe par quelques exemples :

— La fonction $f(x) = \frac{1}{2}\|Ax-b\|^2$ est convexe à gradient Lipschitz de constante $L = \lambda_{max}(A^*A)$. On a en effet $\nabla f(x) = A^*(Ax-b)$. Donc

$$\|\nabla f(x) - \nabla f(y)\| = \|A^*A(x-y)\| \leq \lambda_{max}(A^*A)\|x-y\|.$$

— La fonction $f(x) = -\log(x)$ est convexe sur $\mathbb{R}_+^*$, mais son gradient n'est pas Lipschitz. En effet, $f''(x) = \frac{1}{x^2} \geq 0,\ \forall x > 0$, mais $\lim_{x\to 0^+} f''(x) = +\infty$.

— La fonction $f(x) = \exp(x)$ est convexe sur $\mathbb{R}$, mais son gradient n'est pas Lipschitz. Par contre sur tout intervalle borné du type $[a,b]$, on a $f''(x) \leq \exp(b)$.

1.7.2 Fonctions fortement convexes

L'hypothèse de fonction à gradient Lipschitz ne suffit pas toujours. En particulier, il est impossible de dire quelque chose sur la distance au minimiseur pour des algorithmes de premier ordre sous la simple hypothèse qu'on minimise une fonction à gradient Lispchitz. Une classe de fonctions sur laquelle on peut dire plus de choses est la classe des fonctions fortement convexes.

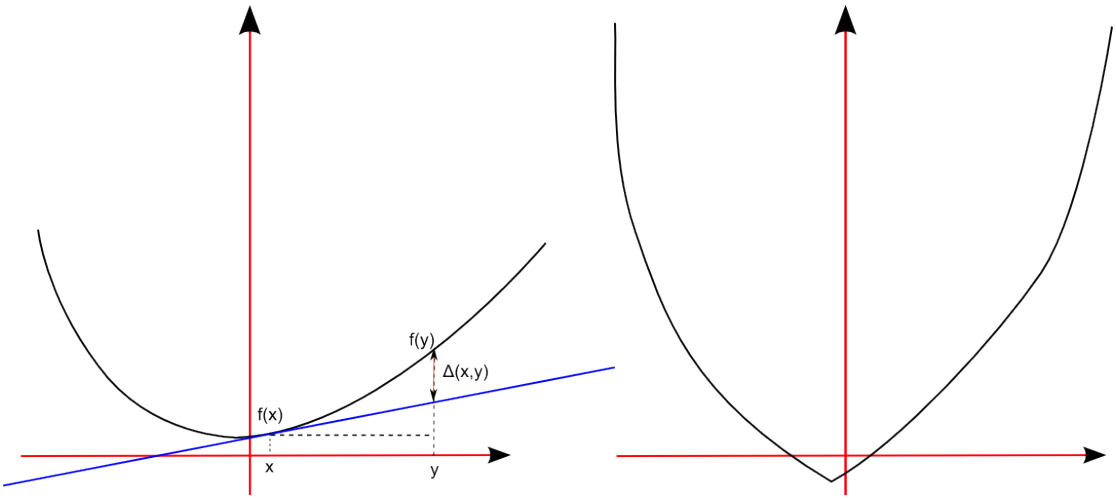


FIGURE 1.11 – Fonctions fortement convexes. Le graphe de la courbe s'éloigne rapidement de la tangente : $\Delta(x,y) \geq \frac{\mu}{2}\|y-x\|_2^2$. Notez qu'une fonction fortement convexe peut être non-différentiable et à domaine borné (voir figure de droite).

Définition 1.14 *Une fonction $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ est dite fortement convexe s'il existe $\mu > 0$ tel que $\forall (x,y) \in \mathbb{R}^n \times \mathbb{R}^n$, $\forall \eta \in \partial f(x)$ on aie :*

$$f(y) \geq f(x) + \langle \eta, y-x \rangle + \frac{\mu}{2}\|y-x\|^2. \tag{1.21}$$

Notez que le signe dans l'inégalité (1.21) est juste inversé par rapport aux fonctions différentiables à gradient Lipschitz. La forte convexité indique donc que le graphe de la courbe s'éloigne suffisamment rapidement de la tangente.

Proposition 1.5 *Une fonction fortement convexe est strictement convexe, elle admet donc un minimiseur unique.*

Note : par contre, une fonction strictement convexe n'est pas forcément fortement convexe (exemple : $f(x) = -\log(x)$).

Preuve. Il suffit de voir que l'équation (1.21) implique que $\forall (x,y) \in \mathbb{R}^n \times \mathbb{R}^n,\ x \neq y$:

$$f(y) > f(x) + \langle \eta, y-x \rangle$$

qui est une inégalité de stricte convexité. □

Les fonctions fortement convexes de classes C^2 sont caractérisées par leur hessienne.

Proposition 1.6 *Une fonction C^2 est μ-fortement convexe si et seulement si*

$$\lambda_{min}(H[f](x)) \geq \mu,\ \forall x \in \mathbb{R}^n.$$

Preuve. Exercice de TD. □

La proposition suivante est un des éléments qui permet d'obtenir des taux de convergence en norme :

Proposition 1.7 *Soit f une fonction μ-fortement convexe qui admet pour minimiseur x^*. Pour tout $x \in \mathbb{R}^n$, on a :*

$$f(x) \geq f(x^*) + \frac{\mu}{2}\|x - x^*\|^2.$$

Preuve. Il suffit d'appliquer l'inégalité (1.21) en $x = x^*$. Ainsi $0 \in \partial f(x)$. □

Proposition 1.8 *Soient f_1 et f_2 deux fonctions convexes de paramètres de forte convexité respectifs $\mu_1 \geq 0$ et $\mu_2 \geq 0$ [a]. Soient α et β deux réels positifs. Alors la fonction $f = \alpha f_1 + \beta f_2$ est μ-fortement convexe avec $\mu \geq \alpha\mu_1 + \beta\mu_2$.*

a. Notez que ces fonctions sont simplement convexes si $\mu_i = 0$.

Preuve. On a $\forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n$:

$$f_1(y) \geq f_1(x) + \langle \eta_1, y - x\rangle + \frac{\mu_1}{2}\|y - x\|^2$$
$$f_2(y) \geq f_2(x) + \langle \eta_2, y - x\rangle + \frac{\mu_2}{2}\|y - x\|^2$$

Il suffit donc de multiplier ces deux inégalités par α et β puis de les additionner. □

Pour finir donnons quelques exemples de fonctions fortement convexes :

- La fonction $f(x) = \frac{1}{2}\|x - x_0\|^2$ est 1 fortement convexe.
- La fonction $f(x) = g(x) + \frac{1}{2}\|x - x_0\|^2$ où g est convexe est 1 fortement convexe.
- La fonction $f(x) = \frac{1}{2}\|Ax - b\|^2$ est fortement convexe si A est de rang plein. Sa constante de forte convexité est alors égale à $\lambda_{min}(A^*A)$.
- La fonction $f(x) = -\log(x)$ n'est pas fortement convexe sur $\mathbb{R}_+^*$. Par contre, elle l'est sur tout intervalles $[a, b]$, $0 < a < b$. Sa constante de forte convexité est alors égale à $\frac{1}{b^2}$.

1.7.3 Conditionnement d'une fonction

Les notions de constante de Lipschitz du gradient et de paramètre de forte convexité généralisent la notion de plus grande et plus petite valeur singulière d'une matrice, comme le montrent les propositions 1.13 et 1.6. Elles permettent de définir une notion de conditionnement d'une fonction.

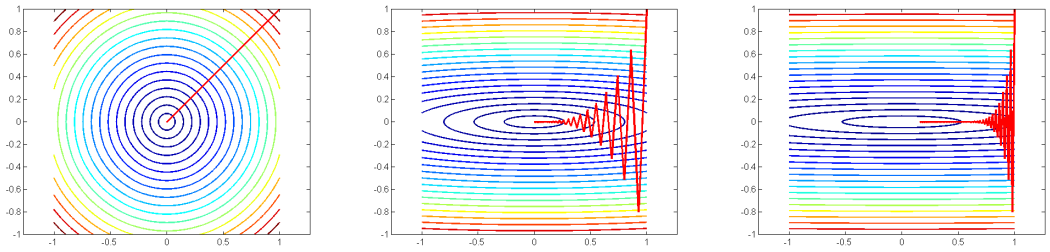


FIGURE 1.12 – Lignes de niveau de la fonction f_α pour $\alpha \in \{1, 5, 10\}$. La courbe rouge indique la trajectoire d'une descente de gradient. Plus le conditionnement α est elevé, plus les courbes de niveau sont allongées et plus la méthode de descente de gradient oscille.

Définition 1.15 *Soit f une fonction μ-fortement convexe, différentiable, telle que ∇f est L-Lipschitz. Le conditionnement de f est défini par la quantité :*

$$Q_f = \frac{L}{\mu}$$

Le conditionnement apparait dans les taux de convergence de toutes les méthodes de premier ordre, de la même façon qu'en algèbre linéaire.

Pour toute fonction f, on a $\mu \leq L$ (pour s'en convaincre comparer les inégalités (1.21) et (1.12) d'où :

$$Q_f \in [1, +\infty]$$

. Le conditionnement est infini dès lors que la fonction f est non différentiable ou non fortement convexe. De façon assez générale, les problèmes mal conditionnés tels que $Q_f >> 1$ sont plus difficile à résoudre que les problèmes bien conditionnés. La figure 1.12 montre les lignes de niveau de la fonction :

$$f_\alpha(x) = \frac{1}{2}\|D_\alpha x\|_2^2$$

avec

$$D_\alpha = \begin{pmatrix} 1 & 0 \\ 0 & \alpha \end{pmatrix}$$

et $\alpha \geq 1$. La hessienne de f_α est D_α et le conditionnement est donc $Q_{f_\alpha} = \alpha$. Plus α est grand, plus les lignes de niveaux sont resserrées. Une méthode de gradient a alors tendance à converger lentement en oscillant dans la "'vallée"' qui mène au minimiseur.

Chapitre 2

Algorithmes pour l'optimisation non différentiable

Dans ce chapitre, nous nous intéressons aux méthodes d'optimisation adaptées au cas où la fonction f à minimiser est convexe mais non différentiable. De manière générale, les méthodes pour l'optimisation non différentiable s'inspirent largement des méthodes différentiables. Pourquoi alors étudier des méthodes spéciales ? L'optimisation différentiable présente des pièges dans lesquels un utilisateur non averti pourrait tomber :

1. **Piège du test d'arrêt.** La condition "$\|g_k\| \leq \varepsilon,\ g_k \in \partial f(x_k)$", traduite directement de la condition "$\nabla f(x_k) \leq \varepsilon$" dans le cas différentiable, peut ne jamais être vérifiée. Prenons par exemple : $f(x) = |x|,\ x \in \mathbb{R}$. Tant que $x_k \neq 0$ (l'optimum global) :
$$\partial f(x_k) = \{1, -1\}.$$
2. **Piège des sous-gradients approchés.** En pratique, le gradient (et même la fonction elle-même) n'est pas toujours calculé exactement : il est souvent obtenu par différences finies. Cette approche n'est plus valable quand le sous-gradient n'est pas continu.
Ex : $f(x) = \begin{cases} x^2 & \text{si } x \geq 0 \\ -x & \text{si } x < 0 \end{cases}$ Par différence finie autour de 0, on obtient :
$$\forall h > 0,\ \frac{f(h) - f(0)}{h} = h > 0,$$
alors que $\partial f(0) = [-1, 0]$. Donc $h \notin \partial f(0)$.
3. **Malédiction du sous-différentiable.** L'application $x \mapsto \partial f(x)$ n'étant pas continue une petite variation de x_k peut entrainer de grandes variations de $\partial f(x_k)$. Basé sur l'information disponible du sous-différentiel, le calcul d'une direction de recherche d_k peut varier énormément et produire des x_{k+1} très différents.

2.1 Un rapide tour d'horizon en optimisation non lisse

Les problèmes d'optimisation non lisse, même en l'absence de contraintes, sont en général très difficiles à résoudre. Il existe actuellement trois grandes familles de méthodes :

les méthodes de *sous-gradients* [22, 25], les méthodes de *faisceaux* [12, 28, 14, 7, 15] et plus récemment les méthodes d'*échantillonnage de gradient* (sous-gradient) [3, 4].

Les méthodes de sous-gradients sont une généralisation naturelle des méthodes de gradient au cas non différentiable en remplaçant le gradient par *un* sous-gradient arbitraire. Ces méthodes (appelées *méthodes de Kiev*) ont été introduites dans les années 60 par N.Z. Shor pour la minimisation sans contrainte de fonctions convexes, et l'on trouve les premiers résultats de convergence dans les travaux de Y.M. Ermoliev [6] ou de B.T. Polyak [23]. Une référence classique sur le sujet est le livre de N.Z. Shor [25]. Le succès de ces méthodes est dû à la simplicité de leur mise en œuvre, en particulier en grande dimension. Bien qu'elles soient largement utilisées en optimisation non lisse, elles souffrent de plusieurs inconvénients : contrairement aux méthodes de gradient en optimisation différentiable, ce ne sont pas des méthodes de descente. Cela peut conduire à des phénomènes oscillatoires et à de mauvaises performances numériques. On ne dispose pas de critère d'arrêt naturel et implémentable permettant de détecter l'optimalité. Enfin, leur vitesse de convergence est en général assez faible (sous-linéaire) même s'il est possible de maintenir une convergence linéaire par des techniques de dilatation en espace le long du gradient [25, Chapitre 3].

Les méthodes de faisceaux ont constitué une avancée majeure pour la résolution de problèmes d'optimisation non différentiable. Historiquement, les méthodes de faisceaux sont basées sur la méthode des plans sécants [5, 9], pour la minimisation sans contrainte de fonctionnelles convexes. Le principe est le suivant : au lieu de faire appel à l'oracle pour générer des candidats à la descente, on se sert des informations de premier ordre de la fonction objectif f pour construire un modèle (convexe) de f, plus facile à minimiser. Etant donné un faisceau d'informations : $\{(x_i, f_i = f(x_i), s_i \in \partial f(x_i))\ ;\ i = 1, \ldots, k\}$ obtenu à partir des itérations précédentes, on construit une approximation convexe linéaire par morceaux de la fonction f (cf Figure 2.1) :

$$\forall y \in \mathbb{R}^n,\ \phi_k(y) = \max_{i=1,\ldots,k} \{f_i + \langle s_i, y - x_i\rangle\} \ (\leq f(y)).$$

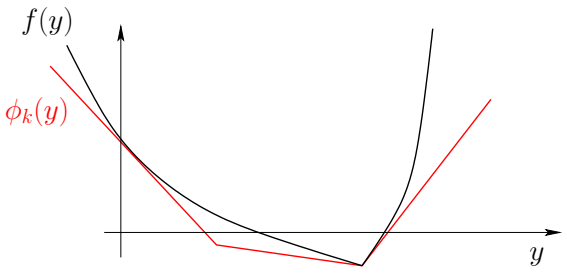


FIGURE 2.1 – Méthode des plans sécants - Construction d'un modèle convexe linéaire par morceaux d'une fonction f convexe, non différentiable.

A chaque itération de l'algorithme, la minimisation du modèle ϕ_k donne un nouvel itéré x_{k+1} qui va permettre d'enrichir le modèle courant par la donnée d'un nouveau plan sécant. L'information passée est ainsi conservée au fil des itérations.

Ces méthodes souffrent de nombreuses limitations : en effet, pour des fonctions convexes très générales, la méthode des plans sécants peut être très instable et avoir de mauvaises performances numériques, [19, Chapitre 4, Section 4.3.6] et [2, Chapitre 9, Exemple 9.7]. Comme pour les algorithmes de sous-gradient, ce ne sont pas des méthodes de descente de la fonction f. Une autre limitation provient de l'accumulation infinie de plans sécants dans le faisceau ce qui augmente à chaque itération la complexité du problème d'optimisation à résoudre, en termes de taille du problème à résoudre mais aussi de conditionnement.

Les méthodes de faisceaux sont des méthodes de stabilisation de la méthode des plans coupants. Considérées en 2001 comme "l'outil le plus efficace et le plus prometteur pour l'optimisation non lisse" [18], elles ont été largement éprouvées et ont fait leurs preuves dans le cas convexe [12, 28, 14, 7, 18, 2]. L'hypothèse de convexité étant un élément clé de ces méthodes, leur généralisation au cas non convexe est loin d'être immédiate et fait l'objet de recherche actuelles.

Depuis 2002, une nouvelle approche basée sur des algorithmes d'échantillonnage de gradient (ou sous-gradients) et développée par J.V. Burke, A.S. Lewis et M.L. Overton [3, 4], s'est peu à peu imposée en méthode concurrente des méthodes de faisceaux, tant sur le plan théorique que sur le plan des performances numériques. Cette méthode, qui peut être vue comme une version stabilisée de l'algorithme de plus profonde descente, est dédiée à la minimisation de fonctions localement lipschitziennes, supposées de classe C^1 sur un sous-ensemble ouvert dense de $\mathbb{R}^n$. La fonction objectif est possiblement non lisse, non convexe. L'idée centrale est d'approcher le sous-différentiel de la fonction objectif par l'enveloppe convexe d'un échantillonnage aléatoire de gradients calculés dans un voisinage du point courant. Un premier résultat de convergence avec une probabilité 1 a été démontré dans [3, 4], puis généralisé par K.C. Kiwiel dans [10]. A noter également une extension de cet algorithme au cas sans dérivée [11].

Une implémentation de ces méthodes est en accès libre [1] sous la forme d'un package Matlab HANSO (Hybrid Algorithm for Non-Smooth Optimization). Sans entrer dans les détails, les algorithmes utilisés dans HANSO mettent en œuvre en première phase un algorithme de type BFGS (Broyden-Fletcher-Goldfarb-Shannon). Cet algorithme, appliqué à des problèmes d'optimisation non lisses, se comporte relativement bien en pratique : la convergence est rapide et conduit souvent à une approximation raisonnablement satisfaisante de l'optimum. Ce phénomène a tout d'abord été observé par plusieurs auteurs : C. Lemaréchal en 1982 [13], L. Lukšan et J. Vlček en 1999 et 2001 [17, 26] avant d'être étudié de façon plus approfondie par A. Lewis et M. Overton [16] très récemment. L'idée est que les fonctions localement lipschitziennes sont différentiables presque partout et qu'en pratique les itérés ne tombent pas sur les points de non-différentiabilité. C'est pour cela que la plupart des codes existants applique en première phase un algorithme de type BFGS, avant d'utiliser des méthodes plus sophistiquées de type faisceaux ou échantillonnage de gradients.

1. http ://www.cs.nyu.edu/faculty/overton/software/hanso

2.2 Méthodes de sous-gradients

Sachant que le sous-différentiel n'est souvent pas connu dans son intégralité, supposons que l'on dispose de l'oracle suivant :

Pour tout $x \in \mathbb{R}^n$, on sait calculer la valeur $f(x)$ de la fonction objectif f en x et au moins un sous-gradient $s \in \partial f(x)$.

Cet oracle fournira dans les algorithmes des candidats à la descente (et non pas nécessairement des directions de descente.

Dans cette partie, on s'intéresse aux méthodes de premier ordre appliquées à la classe de problèmes suivante :

$$\min_{x \in X \subset \mathbb{R}^n} f(x) \tag{2.1}$$

où $X \subset \mathbb{R}^n$ est un ensemble convexe, $f : X \to \mathbb{R}$ est convexe et f est Lipschitz de constante L sur X.

On note x^* un minimiseur du problème (2.1). Notons que ce problème n'a pas de raison particulière d'admettre un minimiseur unique.

2.2.1 Cas sans contrainte

Algorithm 1: Algorithme de sous-gradient dans le cas sans contrainte).

Input:
N : un nombre d'itérations.
$x_0 \in X$: un point de départ.
Output:
x_N : une solution approchée.
begin
  for k *allant de* 0 *à* N **do**
    Calculer $s_k \in \partial f(x_k)$.
    Recherche linéaire : chercher un pas $\tau_k > 0$ tel que :

$$f(x_k - \tau_k \frac{s_k}{\|s_k\|}) < f(x_k).$$

    $x_{k+1} = x_k - \tau_k \frac{s_k}{\|s_k\|}$.

On remarquera que le sous-gradient est ici normalisé, contrairement à la descente de gradient. Il est facile de comprendre l'intérêt de cette normalisation de façon géométrique. Dans le cas différentiable, le gradient tend vers 0 lorsqu'on se rapproche du minimiseur. Dans le cas non différentiable, ce n'est pas le cas. Par exemple, la fonction $f(x) = |x|$ satisfait $\partial f(x) = signe(x)$ pour $x \neq 0$. Le sous-gradient est donc de norme 1 même très près du minimiseur.

On rappelle également que dans le cas non-différentiable, une direction opposée à un sous-gradient n'est pas nécessairement une direction de descente. En conséquence, la recherche linéaire peut échouer ! Bien que l'algorithme 2 ne soit pas de descente, un choix adéquat des τ_k rend malgré tout la méthode convergente.

Théorème 2.1 *Si la suite $(\tau_k)_k$ vérifie :*

$$\lim_{k\to+\infty} \tau_k = 0, \qquad \sum_{k=0}^{+\infty} \tau_k = +\infty,$$

alors : $\liminf f(x_k) = f(x^*)$. *Si, de plus, l'ensemble des minimiseurs de f est borné, alors :*

$$\lim_{k\to+\infty} f(x_k) = f(x^*).$$

Théorème 2.2 *Si la suite $(\tau_k)_k$ vérifie :*

$$\sum_{k=0}^{+\infty} \tau_k = +\infty, \qquad \sum_{k=0}^{+\infty} \tau_k^2 < +\infty,$$

alors la suite $(x_k)_{k\in\mathbb{N}}$ converge vers un minimiseur de f.

Les deux conditions $\lim_{k\to+\infty} h_k = 0$ et $\sum_{k=0}^{\infty} h_k = +\infty$ indiquent que les pas doivent tendre vers 0, mais pas trop rapidement. Par exemple, une décroissance en $O\left(\frac{1}{k}\right)$ est trop rapide car la série $\sum \frac{1}{k^2}$ converge. Par contre une décroissance en $O\left(\frac{1}{\sqrt{k}}\right)$ satisfait les deux conditions. Si la suite $(h_k)_{k\in\mathbb{N}}$ ne tendait pas vers 0, on observerait dans beaucoup de cas des oscillations autour du minimum et pas de convergence. Par exemple, il est facile de voir que la méthode de descente de sous-gradient à pas constant ne converge pas si on l'applique à la fonction $f(x) = |x|$.

2.2.2 Méthodes de sous-gradient projeté

Revenons maintenant au problème :

$$\min_{x\in X} f(x).$$

On note $P_X(x)$ la projection orthogonale du point $x \in \mathbb{R}^n$ sur le domaine admissible X. Comme l'ensemble X est convexe, la projection est univaluée. Notons que la projection est facile à calculer dans le cas de contraintes de type "boite" mais très difficile à calculer dans la plupart des cas.

Algorithm 2: Algorithme de sous-gradient dans le cas sans contrainte).

Input:
N : un nombre d'itérations.
$x_0 \in X$: un point de départ.
Output:
x_N : une solution approchée.
begin
  for k *allant de* 0 *à* N **do**
    Calculer $s_k \in \partial f(x_k)$.
    Recherche linéaire : chercher un pas $\tau_k > 0$ tel que :

$$f(x_k - \tau_k \frac{s_k}{\|s_k\|}) < f(x_k).$$

    $x_{k+1} = p_X(x_k - \tau_k \frac{s_k}{\|s_k\|})$.

Théorème 2.3 *On note* $f_k^* = \min_{k\in\{0,\cdots,k\}} f(x_k)$ *et* R *le diamètre de l'ensemble* X. *Alors :*

$$f_k^* - f^* \leq L\frac{R^2 + \sum_{i=0}^k h_i^2}{2\sum_{i=0}^k h_i}.$$

En particulier si $h_k = \frac{R}{\sqrt{N+1}}$, *on a un taux "optimal" et :*

$$f_k^* - f^* \leq \frac{LR}{\sqrt{N+1}}.$$

Preuve. On commence par montrer que le fait que f soit Lispchitz implique que les sous-gradients de f sont bornés. On a $\forall (x, y) \in X^2$:

$$|f(x) - f(y)| \leq L\|x - y\|$$

or par définition du sous-différentiel :

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

où $\eta \in \partial f(x)$. En combinant ces deux inégalités, on obtient :

$$|\langle \eta, y - x\rangle| \leq L\|x - y\|, \quad \forall y \in X.$$

En particulier pour $y = \eta + x$, on obtient $\|\eta\| \leq L$. Les sous-gradients ont donc une norme majorée par L.

Contrairement aux descentes de gradient, la preuve de convergence ne repose pas ici sur la décroissance monotone de la fonction coût, mais sur le fait que la distance au minimiseur

diminue. On a :

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &= \|x_k - x^* - h_k \frac{\eta_k}{\|\eta_k\|}\|^2 \\ &= \|x_k - x^*\|^2 + h_k^2 - 2h_k\langle x_k - x^*, \frac{\eta_k}{\|\eta_k\|}\rangle.\end{aligned}$$

Or $f(x^*) \geq f(x_k) + \langle \eta_k, x^* - x_k\rangle$ (par définition du sous-gradient). Donc

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + h_k^2 - 2\frac{h_k}{\|\eta_k\|}(f(x_k) - f(x^*)).$$

En sommant de $k = 0$ à N et en utilisant le fait que $\|\eta\| \leq L$ on obtient :

$$\|x_{N+1} - x^*\|^2 \leq \|x_0 - x^*\|^2 + \sum_{k=0}^{N} h_k^2 - \frac{2h_k}{L}(f(x_k) - f(x^*))$$

D'où :

$$\begin{aligned}2\sum_{k=0}^{N} \frac{h_k}{L}(f(x_k) - f(x^*)) &\leq -\|x_{N+1} - x^*\|^2 + \|x_0 - x^*\|^2 + \sum_{k=0}^{N} h_k^2 \\ &\leq R^2 + \sum_{k=0}^{N} h_k^2.\end{aligned}$$

Ce qui permet de conclure :

$$\min_{k\in\{0,\cdots,N\}} f(x_k) - f(x^*) \leq \frac{R^2 + \sum_{k=0}^{N} h_k^2}{2\sum_{k=0}^{N} \frac{h_k}{L}}.$$

□

2.2.3 Application aux problèmes duaux

2.3 Les descentes de gradient proximales

Dans cette partie, on considère le problème suivant :

$$\min_{x\in\mathbb{R}^n} f(x) = g(x) + h(x) \tag{2.2}$$

où :

- $g : \mathbb{R}^n \to \mathbb{R}$ est une fonction convexe, différentiable de gradient L Lipschitz :

$$\|\nabla g(x_1) - \nabla g(x_2)\| \leq L\|x_1 - x_2\|,\ \forall (x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n. \tag{2.3}$$

- $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ est une fonction convexe, fermée.

On suppose que ce problème admet au moins une solution et on note X^* l'ensemble (convexe fermé) des minimiseurs. Les conditions d'optimalité de ce problème sont :

$$0 \in \partial h(x) + \nabla g(x) \Leftrightarrow -\nabla g(x) \in \partial h(x).$$

2.3.1 Opérateurs proximaux

Définition 2.1 (Opérateur proximal) *L'opérateur proximal de f est noté* Prox_f *et il est défini pour tout $x \in \mathbb{R}^n$ par :*

$$\mathrm{Prox}_f(x) = \underset{u\in\mathbb{R}^n}{\arg\min}\, f(u) + \frac{1}{2}\|u - x\|_2^2. \tag{2.4}$$

Exemple 2.3.1 *Voici quelques exemples d'opérateurs proximaux.*

— *Si* $f(x) = 0$, $\mathrm{Prox}_f(x) = x$.

— *Si* $f(x) = \chi_X(x) = \begin{cases} 0 & si\ x \in X \\ +\infty & sinon, \end{cases}$ *où $X \subseteq \mathbb{R}^n$ est un ensemble convexe fermé, alors* $\mathrm{Prox}_f(x)$ *est la projection euclidienne de x sur X. En effet :*

$$\mathrm{Prox}_f(x) = \underset{u\in X}{\arg\min}\, \|u - x\|_2^2 = P_X(x). \tag{2.5}$$

— *Si* $f(x) = \alpha\|x\|_1$, Prox_f *est un seuillage doux de x :*

$$(\mathrm{Prox}_f(x))_i = \begin{cases} x_i - \alpha & si\ x_i \geq \alpha \\ 0 & si\ |x_i| \leq \alpha \\ x_i + \alpha & si\ x_i \leq -\alpha. \end{cases} \tag{2.6}$$

Montrons quelques propriétés utiles de cet opérateur.

Proposition 2.1 *Si $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ une fonction convexe fermée,* $\mathrm{Prox}_f(x)$ *existe et est unique pour tout $x \in \mathbb{R}^n$. De plus il est caractérisé par les équations suivantes :*

$$\begin{aligned} & u = \mathrm{Prox}_f(x) \\ \Leftrightarrow\ & 0 \in \partial f(u) + u - x \\ \Leftrightarrow\ & u = (I_n + \partial f)^{-1}(x). \end{aligned}$$

Proposition 2.2 (Contraction et expansivité) *Soit $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ une fonction convexe fermée. Alors* Prox_f *est un opérateur fortement non expansif :*

$$\langle \mathrm{Prox}_f(x_1) - \mathrm{Prox}_f(x_2), x_1 - x_2 \rangle \geq \|\mathrm{Prox}_f(x_1) - \mathrm{Prox}_f(x_2)\|_2^2. \tag{2.7}$$

Il est Lipschitz continu de constante 1 :

$$\|\mathrm{Prox}_f(x_1) - \mathrm{Prox}_f(x_2)\|_2 \leq \|x_1 - x_2\|_2. \tag{2.8}$$

Preuve. On a :

$$x_1 - \mathrm{Prox}_f(x_1) \in \partial f(\mathrm{Prox}_f(x_1)) \tag{2.9}$$

et

$$x_2 - \mathrm{Prox}_f(x_2) \in \partial f(\mathrm{Prox}_f(x_2)) \tag{2.10}$$

or si f est convexe fermée, l'opérateur multivalué ∂f est monotone, ce qui signifie que :

$$\langle \eta_1 - \eta_2, u_1 - u_2 \rangle \geq 0, \ \forall \eta_1 \in \partial f(\mathrm{Prox}_f(x_1)), \forall \eta_2 \in \partial f(\mathrm{Prox}_f(x_2)). \tag{2.11}$$

(Cette inégalité permet d'exprimer le fait que dans le cas C^2, la hessienne d'une fonction convexe est semi-définie positive). L'inégalité (2.8) est obtenue en appliquant le théorème de Cauchy-Schwartz. □

Proposition 2.3 (Identité de Moreau) *Soit $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ une fonction convexe fermée. On a alors :*

$$x = \mathrm{Prox}_f(x) + \mathrm{Prox}_{f^*}(x). \tag{2.12}$$

Preuve. On note $u = \mathrm{Prox}_f(x)$ et $v = x - u$. Le vecteur u satisfait $0 \in \partial f(u) + x - u$ soit encore $v \in \partial f(u)$. On a donc $u \in \partial f^*(v)$ soit encore $x - v \in \partial f^*(v)$ ou $v = \mathrm{Prox}_{f^*}(x)$. □

Exemple 2.3.2 *Soit L un sous-espace vectoriel de $\mathbb{R}^n$ et $L^\perp$ son complément orthogonal. La propriété précédente n'est alors rien d'autre que la décomposition orthogonale :*

$$x = P_L(x) + P_{L^\perp}(x). \tag{2.13}$$

2.3.2 Descente de gradient proximale

Définition 2.2 (Descente de gradient proximale)

$$x_{k+1} = \mathrm{Prox}_{th}(x_k - t\nabla g(x_k)). \tag{2.14}$$

Malheureusement, F n'est pas contractante en général, mais seulement non expansive. Les preuves de convergence linéaires qui reposent sur des itérées de contractions ne fonctionnent donc pas pour analyser ce schéma. De façon générale, on obtient tout de même des résultats de convergence sous-linéaire :

Théorème 2.4 *Le schéma de descente* (2.14) *assure que $f(x_k) - f(x^*) \leq \frac{L\|x_0 - x^*\|_2^2}{k}$ pour $t = \frac{1}{L}$.*

Pour démontrer ce théorème on commence par réécrire l'itération (2.14) sous la forme :

$$x_{k+1} = x_k - tG_t(x) \tag{2.15}$$

avec

$$G_t(x) = \frac{1}{t}\left(x - Prox_{th}(x - t\nabla g(x))\right). \tag{2.16}$$

La fonction $G_t(x)$ peut être interprétée comme une direction de descente au point x. Notons que $G_t(x) = 0 \Leftrightarrow x \in X^*$ et que

$$G_t(x) \in \nabla g(x) + \partial h(x - tG_t(x)). \tag{2.17}$$

On a aussi besoin des lemmes suivants :

Lemme 2.1 *La fonction g satisfait les inégalités suivantes :*

$$g(x_2) \geq g(x_1) + \langle \nabla g(x_1), x_2 - x_1 \rangle \quad \textit{(convexité)} \tag{2.18}$$

$$g(x_2) \leq g(x_1) + \langle \nabla g(x_1), x_2 - x_1 \rangle + \frac{L}{2}\|x_2 - x_1\|_2^2 \quad \textit{(gradient Lipschitz).} \tag{2.19}$$

Lemme 2.2 (Inégalité globale) *Pour $t \in]0, \frac{1}{L}]$, et pour tout $(x, z) \in \mathbb{R}^n \times \mathbb{R}^n$ on a :*

$$f(x - tG_t(x)) \leq f(z) + \langle G_t(x), x - z) \rangle - \frac{t}{2}\|G_t(x)\|_2^2. \tag{2.20}$$

Preuve. On pose $v = G_t(x) - \nabla g(x)$. On a :

$$\begin{aligned}
& f(x - G_t(x)) \\
&\leq g(x) - t\langle \nabla g(x), G_t(x) \rangle + \frac{t}{2}\|G_t(x)\|_2^2 + h(x - tG_t(x)) \\
&\leq g(z) + \langle \nabla g(x), x - z \rangle - t\langle \nabla g(t), G_t(x) \rangle + \frac{t}{2}\|G_t(x)\|_2^2 \\
&\quad + h(z) + \langle v, x - z - tG_t(x) \rangle \\
&\leq g(z) + h(z) + \langle G_t(x), x - z \rangle - \frac{t}{2}\|G_t(x)\|_2^2.
\end{aligned}$$

Le passage de la première ligne à la deuxième est liée à la convexité de g et de h at au fait que $v \in \partial h(x - tG_t(x))$. □ Ce lemme est important car en notant $x^+ = x - tG_t(x)$:

— en prenant $z = x$, on voit que la méthode est une méthode de descente :

$$f(x^+) \leq f(x) - \frac{t}{2}\|G_t(x)\|_2^2. \tag{2.21}$$

— en prenant $z = x^*$, on voit que la distance au minimiseur décroît [2] :

$$\begin{aligned}
0 \leq\ & f(x^+) - f^* && (2.22)\\
\leq\ & \langle G_t(x), x - x^* \rangle - \frac{t}{2}\|G_t(x)\|_2^2 && (2.23)\\
=\ & \frac{1}{2t}\left(\|x - x^*\|_2^2 - \|x - x^* - tG_t(x)\|_2^2\right) && (2.24)\\
=\ & \frac{1}{2t}\left(\|x - x^*\|_2^2 - \|x^+ - x^*\|_2^2\right). && (2.25)
\end{aligned}$$

Pour conclure, il suffit de poser $x^+ = x_{k+1}$ et $x = x_k$ dans l'inégalité (2.25) et de sommer

2. cette étape de démonstration est très commune pour démontrer des taux de convergence sous-linéaires.

les inégalités de 1 à k :

$$\begin{aligned} 0 &\leq \sum_{i=1^k} f(x_i) - f^* \\ &\leq \frac{1}{2t} \sum_{i=1}^{k} \left(\|x_{i-1} - x^*\|_2^2 - \|x_i - x^*\|_2^2 \right) \\ &\leq \frac{1}{2t} (\|x_0 - x^*\|_2^2 - \|x_k - x^*\|_2^2) \\ &\leq \frac{1}{2t} \|x_0 - x^*\|_2^2. \end{aligned}$$

Comme la suite $(f(x_k))$ est décroissante, on obtient finalement :

$$f(x_k) - f(x^*) \leq \frac{1}{k} \sum_{i=1}^{k} f(x_i) - f(x^*) \leq \frac{1}{2kt} \|x_0 - x^*\|. \tag{2.26}$$

2.4 Dualité pour les problèmes fortement convexes

Dans cette partie, nous nous intéressons au problème suivant :

$$\min_{x \in \mathbb{R}^n} P(x) = g(Ax) + h(x) \tag{2.27}$$

où :

- $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ est une fonction convexe fermée.
- $A \in R^{m \times n}$ est une application linéaire.
- $h : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ est une fonction fortement convexe, de paramètre de forte convexité σ.

L'algorithme présenté précédemment ne peut pas être appliqué en général car aucune des fonctions $g \circ A$ et g ne sont supposées différentiables. De plus, même si h l'était, il est en général impossible d'obtenir une formule explicite de l'application $\text{Prox}_{g \circ A}$. Pour introduire le prochain algorithme, nous présentons d'abord deux résultats important d'analyse convexe.

Théorème 2.5 (Dualité de Fenchel-Rockafellar) *Soient f et g deux fonctions convexes fermées et $A \in \mathbb{R}^m \times n$ une transformée linéaire telles que*

$$A\,(ri)(dom(h)) \cap (ri)(dom(g)) \neq \emptyset. \tag{2.28}$$

Alors :

$$\min_{x \in \mathbb{R}^n} g(Ax) + h(x) = - \inf_{y \in \mathbb{R}^m} g^*(y) + h^*(-A^* y). \tag{2.29}$$

Le problème de gauche est appelé problème primal et celui de droite est appelé problème dual.

Les conditions d'optimalité du problème primal-dual (2.29) sont :

Théorème 2.6

$$y^* \in \partial g(Ax^*) \tag{2.30}$$

$$x^* \in \partial h^*(-A^*y^*). \tag{2.31}$$

Pour finir donnons un résultat :

Théorème 2.7 *Soit $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ une fonction convexe fermée. Les deux propositions suivantes sont équivalentes :*

- *f est différentiable à gradient L-Lipschitz.*
- *f^* est fortement convexe de module $\frac{1}{L}$.*

Ce théorème et l'inclusion (2.31) motive l'utilisation de l'algorithme suivant pour résoudre (2.27) :

Définition 2.3 (Descente de gradient duale) *Générer une suite minimisante (y_k) (par exemple les descentes proximales ou les descentes proximales accélérées) pour le problème suivant :*

$$\min_{y \in \mathbb{R}^m} D(y) = g^*(y) + h^*(-A^*y). \tag{2.32}$$

Définir la suite :

$$x_k = \nabla h^*(-A^*y_k^*). \tag{2.33}$$

Le théorème suivant permet d'obtenir des garanties théoriques sur la rapidité de convergence :

Théorème 2.8 *L'algorithme précédent assure que :*

$$\|x_k - x^*\|_2^2 \leq \frac{2}{\sigma}(d(y_k) - d(y^*)). \tag{2.34}$$

En particulier, si une descente de gradient proximale accélérée est utilisée pour minimiser D, on obtient :

$$\|x_k - x^*\|_2^2 \leq \frac{2\|A\|^2\|y_0 - y^*\|_2^2}{\sigma^2(k+1)^2}. \tag{2.35}$$

Bibliographie

[1] H. H. Bauschke and P. L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. Springer Science & Business Media, 2011.

[2] J. Bonnans, J. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Numerical optimization*. Universitext. Springer-Verlag, Berlin, second edition, 2006. Theoretical and practical aspects.

[3] J. Burke, A. Lewis, and M. Overton. Approximating subdifferentials by random sampling of gradients. *Mathematics of Operations Research*, 27(3) :567–584, 2002.

[4] J. Burke, A. Lewis, and M. Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM Journal on Optimization*, 15(3) :751–779, 2005.

[5] E. W. Cheney and A. A. Goldstein. Newton's method for convex programming and Tchebycheff approximation. *Numerische Mathematik*, 1 :253–268, 1959.

[6] Y. Ermoliev. Stochastic programming methods. *Nauka, Moscow*, 7(136), 1976.

[7] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms. II*, volume 306 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, 1993. Advanced theory and bundle methods.

[8] J.-B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms I : fundamentals*, volume 305. Springer Science & Business Media, 2013.

[9] J. E. Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8 :703–712, 1960.

[10] K. Kiwiel. Convergence of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 18(2) :379–388, 2007.

[11] K. Kiwiel. A nonderivative version of the gradient sampling algorithm for nonsmooth nonconvex optimization. *SIAM Journal on Optimization*, 20(4) :1983–1994, 2010.

[12] K. C. Kiwiel. *Methods of descent for nondifferentiable optimization*, volume 1133 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1985.

[13] C. Lemaréchal. Numerical experiments in nonsmooth optimization. In E. e. Nurminski, editor, *Progress in Nondifferentiable Optimization*, pages 61–84. International Institute for Applied Systems Analysis, 1982.

[14] C. Lemaréchal. Chapter VII Nondifferentiable optimization. In *Optimization*, volume 1 of *Handbooks in Operations Research and Management Science*, pages 529–572. Elsevier, 1989.

[15] C. Lemaréchal and F. Oustry. Nonsmooth Algorithms to Solve Semidefinite Programs. In L. El Ghaoui and S.-I. Niculescu, editors, *Advances in Linear Matrix Inequality Methods in Control*, chapter 3. SIAM, 2000.

[16] A. Lewis and M. Overton. Nonsmooth optimization via quasi-newton methods. *Mathematical Programming*, 141(1) :135–163, 2013.

[17] L. Lukšan and J. Vlček. Globally convergent variable metric method for convex nonsmooth unconstrained minimization. *Journal of Optimization Theory and Applications*, 102(3) :593–613, 1999.

[18] M. M. Mäkelä. Survey of bundle methods for nonsmooth optimization. *Optim. Methods Softw.*, 17(1) :1–29, 2002.

[19] A. Nemirovskii and D. Yudin. Problem complexity and method efficiency in optimization. In *Discrete Mathematics (Original Russian : Nauka 1979)*. Wiley-Intersciences Series, 1983.

[20] Y. Nesterov. *Introductory lectures on convex optimization*, volume 87. Springer Science & Business Media, 2004.

[21] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1) :125–161, 2013.

[22] B. Polyak. Subgradient methods : A survey of soviet research. In *Nonsmooth optimization : Proceedings of the IIASA workshop March*, pages 5–30, 1977.

[23] B. Polyak. Subgradient methods : a survey of Soviet research. In *Nonsmooth optimization (Proc. IIASA Workshop, Laxenburg, 1977)*, volume 3 of *IIASA Proc. Ser.*, pages 5–29. Pergamon, Oxford-New York, 1978.

[24] R. T. Rockafellar. *Convex analysis*. Princeton university press, 2015.

[25] N. Z. Shor. *Minimization methods for nondifferentiable functions*, volume 3 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1985. Translated from the Russian by K. C. Kiwiel and A. Ruszczyński.

[26] J. Vlček and L. Lukšan. Globally convergent variable metric method for nonconvex nondifferentiable unconstrained minimization. *Journal of Optimization Theory and Applications*, 111(2) :407–430, 2001.

[27] C. Zalinescu. *Convex analysis in general vector spaces*. World Scientific, 2002.

[28] J. Zowe. Nondifferentiable optimization. In *Computational mathematical programming (Bad Windsheim, 1984)*, volume 15 of *NATO Advanced Science Institutes Series F : Computer and Systems Sciences*, pages 323–356. Springer, Berlin, 1984.