

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES



Mémoire présenté en vue de l'obtention du Diplôme :

MASTER en Mathématiques

Option : **Analyse**

Par

Gagui Assia

Titre

Méthodes d'optimisation non différentiable

Membres du Comité d'Examen :

Dr. **Guidad Daradji** UMKB Président

Dr. **Kaci Fatma** UMKB Encadreur

Dr. **Soltani Siham** UMKB Examineur

Juin 2019

DÉDICACE

Je dédie ce travail à :

Ma chère maman pour tout sa tendresse, amour et affection

Mes chers frères et mes soeurs

Toute ma famille et mes amies

Tous mes formateurs pour leurs efforts et leurs amabilités et tous qui m'ont aidé de pré ou de loin pour la réalisation de ce travail.

REMERCIEMENTS

Je tient tout d'abord à remercier «Dieu» le tout puissant et miséricordieux, qui m' a donné la force et la patience d'accomplir ce travail.

Avec une profonde reconnaissance et considération particulière que je remercie mon encadreur «Kaci Fatma» pour la sollicitude avec laquelle elle a suivi et guidé ce travail.

Un grand merci également pour les membres du jury

Je présente tous mes remerciement aux enseignants du département, ainsi que tous ceux qui m'ont soutenue et m'ont aidé tout le long de cette étude et toutes les personnes qui ont contribué directement ou indirectement à ce travail.

Table des matières

| | |
|---|------------|
| Remerciements | ii |
| Table des matières | iii |
| Table des figures | v |
| Introduction | 1 |
| 1 Motivations et Notations Principales | 3 |
| 1.1 Introduction | 3 |
| 1.2 Problème d'optimisation | 4 |
| 1.2.1 Élément de convexité | 4 |
| 1.2.2 Problème d'optimisation | 5 |
| 1.3 Rappel de calcul différentiable | 6 |
| 1.3.1 Dérivée partielle | 6 |
| 1.3.2 Fonction différentiable | 6 |
| 1.3.3 Gradient | 7 |
| 1.3.4 Dérivée directionnelle | 7 |
| 1.3.5 Hessienne | 8 |

| | | |
|----------|--|-----------|
| 1.3.6 | Fonction de classe C^1 | 8 |
| 1.4 | Les minimums locaux et globaux | 9 |
| 1.5 | Existence et unicité d'un point de minimum | 10 |
| 1.5.1 | Remarques sur l'existence et l'unicité | 12 |
| 2 | Quelques méthodes d'optimisation non différentiable | 14 |
| 2.1 | Introduction | 14 |
| 2.2 | Sous-gradient et sous différentielle | 15 |
| 2.2.1 | Règles de calcul de sous-différentiel | 15 |
| 2.3 | Méthode de sous-gradient | 16 |
| 2.3.1 | Direction de descente | 17 |
| 2.3.2 | Condition d'optimalité | 17 |
| 2.3.3 | Algorithme de sous-gradient | 17 |
| 2.3.4 | Méthodes de sous-gradient projeté | 20 |
| 2.3.5 | Quelques résultats de la convergence | 21 |
| 2.3.6 | Choix de polyak | 24 |
| 2.4 | Méthodes des plans sécants et de faisceaux | 25 |
| 2.5 | Exemples | 26 |
| 2.5.1 | Méthode de sous-gradient | 26 |
| 2.5.2 | Méthode sous-gradient projeté | 28 |
| | Bibliographie | 30 |
| | Annexe : Abréviations et Notations | 31 |

Table des figures

| | | |
|-----|---|----|
| 1.1 | En haut : quelques exemples d'ensembles convexes . En bas : quelques exemples d'ensembles non convexes | 4 |
| 1.2 | Un exemple de fonction convexe | 5 |
| 1.3 | Un exemple de fonction coersive $J(x) = \ x \ $ | 9 |
| 1.4 | Minimum et maximum locaux et globaux de la fonction $x \rightarrow 3e^{-x^2} + e^{-(x-3)^2}$ | 10 |
| 2.1 | Méthode de plans sécants -construction d'un modèle convexe linéaire par morceaux d'une fonction convexe non différentiable | 26 |
| 2.2 | la valeur de $f_{best} - f^*$ en fonction de nombre d'itérations k par la méthode du sous-gradient avec une longueur de pas constante τ | 27 |
| 2.3 | La valeur de $f_{best}^{(k)} - f^*$ en fonction du nombre d'itérations k , pour le sous-gradient avec le choix de Polyak (ligne noire continue) et les méthodes de sous-gradient avec une taille de pas décroissante. | 28 |
| 2.4 | la valeur de $f_{best}^{(k)} - f^*$ par rapport au nombre d'itérations k , pour la méthode du sous-gradient avec le choix de Polyak $\tau_k = \frac{100}{k}$ | 29 |

Introduction

L'optimisation est une branche des mathématiques, cherchant à analyser et à résoudre analytiquement ou numériquement les problèmes qui consistent à déterminer le meilleur élément d'un ensemble au sens d'un critère quantitatif donné. Ce mot vient du latin optimum qui signifie le meilleur.

L'optimisation recouvre l'étude des critères d'optimalité pour les différents problèmes, la détermination des méthodes algorithmiques pour résoudre ces problèmes, l'étude de la structure et l'expérimentation de ces méthodes avec des problèmes de la vie courante.

On s'intéresse dans ce mémoire au problème suivant :

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

où $J : C \subset \mathbb{R}^n \rightarrow \mathbb{R}$

Parmi les méthodes les plus utilisées pour résoudre des problèmes de ce type, on peut citer la méthode du sous-gradient. Cette méthode est apparue pour résoudre les problèmes de minimisation d'une fonction non continuellement différentiable. Pour les mêmes problèmes, on a aussi la méthode de sous-gradient projeté et la méthode des plans sécants et de faisceaux.

Le contenu de ce mémoire est divisé en deux chapitres :

Le premier chapitre est surtout un chapitre de généralités qui permet de définir l'ensemble de notations nous aurons besoin dans la suite. Nous présentons les principales définitions de la

convexité, les calculs différentiels ainsi que quelques théorèmes caractérisant l'optimum d'un problème.

Le deuxième chapitre est consacré à la présentation de différentes méthodes d'optimisation pour les problèmes non différentiables en se basant surtout sur la méthode de sous-gradient.

Chapitre 1

Motivations et Notations Principales

1.1 Introduction

Considérons le problème d'optimisation suivant :

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

Où J est une fonction de $\mathbb{R}^n \rightarrow \mathbb{R}$.

Ce chapitre est consacré aux définitions et aux notations principales de l'optimisation, on décrit dans la notation de la différentiabilité d'une fonction , le gradient, la hessienne et quelques éléments de convexité. On présente les résultats d'existence et d'unicité de la solution du problème.

1.2 Problème d'optimisation

1.2.1 Élément de convexité

On note par V un espace vectoriel normé, x un point de V et $J : V \rightarrow \mathbb{R}$

$$\langle x, y \rangle = \sum_i x_i y_i, \quad \|x\| = \left(\sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}}.$$

Définition 1.2.1 (*Ensemble convexe*)

Une partie $C \subset V$ est convexe si : $\forall (x, y) \in C^2, \forall \lambda \in [0, 1],$

$$\lambda x + (1 - \lambda) y \in C.$$



FIG. 1.1 – En haut : quelques exemples d'ensembles convexes . En bas : quelques exemples d'ensembles non convexes

Définition 1.2.2 (*fonction convexe*)

Soit C un convexe d'un espace vectoriel V

La fonction $J : C \rightarrow \mathbb{R}$ est dite convexe si : $\forall (x, y) \in C^2 \forall \lambda \in [0, 1],$

$$J(\lambda x + (1 - \lambda) y) \leq \lambda J(x) + (1 - \lambda) J(y).$$

On dit que J est strictement convexe si :

$$J(\lambda x + (1 - \lambda) y) < \lambda J(x) + (1 - \lambda) J(y).$$

On dit que J est concave si $(-J)$ est convexe.

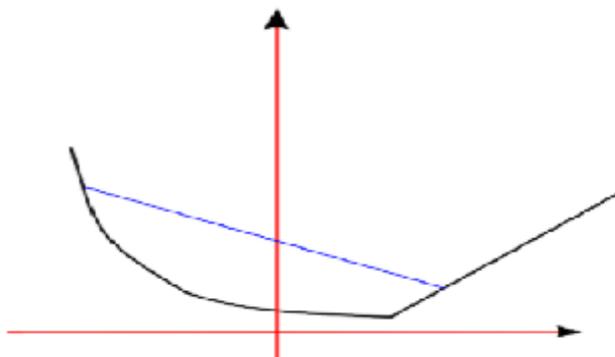


FIG. 1.2 – Un exemple de fonction convexe

1.2.2 Problème d'optimisation

On appelle problème d'optimisation tout problème de la forme :

Trouver x^* tel que :

$$x^* \in U \text{ et } J(x^*) = \min_{x \in U} J(x). \quad (1.1)$$

ou équivalent à $J(x^*) \leq J(x), \forall x \in U$

où $J : U \subset \mathbb{R}^n \rightarrow \mathbb{R}$ est une fonction donnée que l'on appelle fonctionnelle du problème d'optimisation.

- Le but de l'optimisation est de proposer des algorithmes permettant d'approcher les solutions x^* .
- Le problème est dit sans contraintes si $U = \mathbb{R}^n$, et sous contrainte sinon.
- On dit que le problème est convexe si J et U sont convexes.

- Si U est dénombrable : le problème d'optimisation est dit discret.
- Si U est fini : le problème d'optimisation est dit combinatoire
- Si $U \in \mathbb{Z}^n$: le problème d'optimisation est en nombre entier.

1.3 Rappel de calcul différentiable

Soit U un ouvert de \mathbb{R}^n et $J : U \rightarrow \mathbb{R}$

1.3.1 Dérivée partielle

La dérivée partielle de J en x par rapport à x_i , si elle existe, est :

$$J_{x_i}(x) = \frac{\partial J(x)}{\partial x_i} = \lim_{s \rightarrow 0} \frac{J(x_1, \dots, x_i + s, \dots, x_n) - J(x_1, \dots, x_i, \dots, x_n)}{s}.$$

1.3.2 Fonction différentiable

Définition 1.3.1 Soit $x_0 \in U$. On dit que J est différentiable en x_0 s'il existe une application linéaire $L : \mathbb{R}^n \rightarrow \mathbb{R}$ tel que au voisinage de x_0 l'on ait

$$J(x_0 + h) - J(x_0) = Lh + o(|h|).$$

On la note par $DJ(x_0)$.

1.3.3 Gradient

Si J est différentiable, alors on définit le gradient de J en x par :

$$\begin{aligned} \nabla J(x) &= g(x) \\ g(x) &: \mathbb{R}^n \rightarrow \mathbb{R} \end{aligned}$$

Si toutes les dérivées partielles existent alors :

$$g(x) = \left[\frac{\partial J(x)}{\partial x_i} \right]_{i=1, \dots, n} = \begin{bmatrix} \frac{\partial J(x)}{\partial x_1} \\ \cdot \\ \cdot \\ \cdot \\ \frac{\partial J(x)}{\partial x_n} \end{bmatrix}$$

Par exemple, le gradient de la fonction $J(x, y, z) = x^3y - yz + z^2$ est égale

$$\nabla J(x) = (3x^2y, x^3 - z, -y + 2z).$$

1.3.4 Dérivée directionnelle

Définition 1.3.2 La dérivée directionnelle de J en x dans la direction $d \in \mathbb{R}^n$ est :

$$J_d(x) = \lim_{s \rightarrow 0} \frac{J(x + sd) - J(x)}{s} = J'(x, d).$$

Ou encore la dérivée directionnelle est le produit scalaire de la direction avec le gradient

$$\iff J_d(x) = g(x)^T \cdot d$$

1.3.5 Hessienne

Définition 1.3.3 La hessienne de J en x est la matrice carrée, noté $H(x)$:

$$H(x) = \nabla^2 J(x)$$

$$H(x) : \mathbb{R}^n \rightarrow \mathbb{R}$$

$$\begin{aligned} H(x) &= \left[\frac{\partial^2 J(x)}{\partial x_i \partial x_j} \right]_{i,j=1,\dots,n} \\ &= \begin{bmatrix} \frac{\partial^2 J(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 J(x)}{\partial x_1 \partial x_n} \\ \cdots & \cdots & \cdots \\ \frac{\partial^2 J(x)}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 J(x)}{\partial x_n^2} \end{bmatrix} \end{aligned}$$

1.3.6 Fonction de classe C^1

- Soit U un ouvert de \mathbb{R}^n , $J : U \rightarrow \mathbb{R}$. Si toutes les dérivées partielles de J existent et sont continues sur U on dit que J est de classe C^1 sur U , et on écrit $J \in C^1(U)$
- Soit U un ouvert de \mathbb{R}^n , $J : U \rightarrow \mathbb{R}$ et $x_0 \in U$
- Si J est différentiable en x_0 alors :
 - J est continue en x_0
 - J admet toute dérivée directionnelle en x_0 et sa différentielle est donnée par :

$$DJ(x) : \mathbb{R}^n \rightarrow \mathbb{R}$$

$$: h = (h_1, \dots, h_n) \mapsto \partial_{x_1} J(x_1) h_1 + \dots + \partial_{x_n} J(x_n) h_n = \nabla J(x_0) \cdot h$$

Théorème 1.3.1 (Condition suffisante pour la différentiabilité)

Soient U un ouvert de \mathbb{R}^n , $J : U \rightarrow \mathbb{R}$ et $x_0 \in U$. Si J est de classe C^1 au voisinage de x_0 alors elle est différentiable au point x_0 .

Définition 1.3.4 (*fonction coersive*)

Soit $U \subset \mathbb{R}^n$ un ensemble non borné et $J : U \rightarrow \mathbb{R}$.

On dit que J est coersive si on a : $\lim_{x \in U, \|x\| \rightarrow +\infty} J(x) = +\infty$.

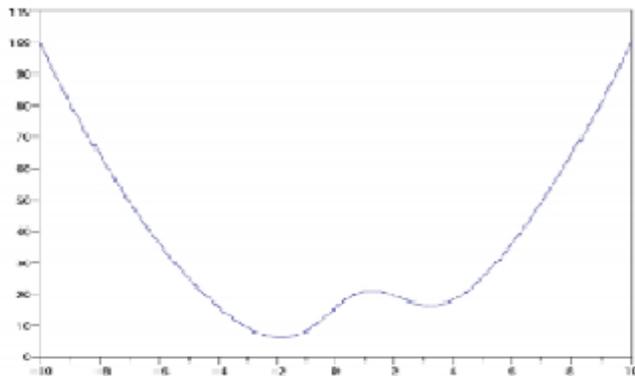


FIG. 1.3 – Un exemple de fonction coersive $J(x) = \| x \|$

1.4 Les minimums locaux et globaux

Définition 1.4.1 (*Minimum global*)

Soit $J : U \subset \mathbb{R}^n \rightarrow \mathbb{R}$ une fonction.

$J(x^*)$ est le minimum global de J si $J(x^*) \leq J(x), \forall x \in U$. On dit que x^* est un minimiseur global de J .

Définition 1.4.2 (*Minimum local*)

Soit $J : U \subset \mathbb{R}^n \rightarrow \mathbb{R}$, $J(x^*)$ est le minimum local de J si $\exists \varepsilon > 0, \forall x \in]x^* - \varepsilon, x^* + \varepsilon[, J(x^*) \leq J(x)$. x^* est un minimiseur local de J .

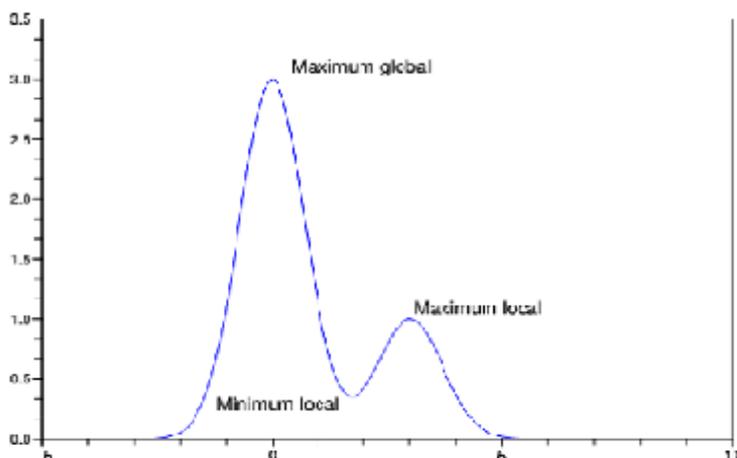


FIG. 1.4 – Minimum et maximum locaux et globaux de la fonction $x \rightarrow 3e^{-x^2} + e^{-(x-3)^2}$

1.5 Existence et unicité d'un point de minimum

Théorème 1.5.1 (*de Weierstrass*)

Si une fonction $J : \mathbb{R}^d \rightarrow \mathbb{R}$ est continue sur un ensemble fermé et borné K (i.e un ensemble compact) alors il existe un point x^* qui atteint le minimum :

$$x^* \in \arg \min_{x \in K} J(x)$$

Théorème 1.5.2 (*Existence*)

Soit $U \subset \mathbb{R}^n$ un ensemble non vide et fermé et $J : U \rightarrow \mathbb{R}$ une fonction continue. On suppose

- soit U est borné.
- soit U est non borné et J est une fonction coercive sur U .

Alors il existe au moins un point de minimum de J sur U (i.e, $\exists x^* \in U$ tel que $J(x^*) \leq J(x), \forall x \in U$)

Preuve. On distingue deux cas :

Cas 1 : l'ensemble U est borné

Alors comme U est aussi fermé, U est compact, comme J est continue, le théorème de

Weierstrass nous assure que J est bornée sur U et elle atteint ses bornes.

Donc il existe au moins un point de minimum *absolu* (*global*) de J sur U

Cas 2 : l'ensemble U est non borné

soit $a \in U$ et considérons l'ensemble :

$$E = \{y \in U, J(y) \leq J(a)\}; a \in E$$

On a E est fermé car : $E = J^{-1}([-\infty, J(a)])$ donc E est l'image inverse d'un intervalle fermé par une fonction continue, de plus il est facile de montrer que E est borné. Supposons le contraire : alors il existe une suite $y_k \in E$ avec $\|y_k\| \rightarrow +\infty$ pour $k \rightarrow +\infty$, comme J est coercive sur U , ceci entraîne : $J(y_k) \rightarrow +\infty$ ce qui est absurde, car $J(y_k) \leq J(a) \forall k \in \mathbb{N}$

On déduit alors que E est un ensemble compact dans \mathbb{R}^n . Du théorème de *Weierstrass*,

$$\exists x^* \in E \text{ tel que } J(x^*) \leq J(x), \forall x \in E \quad \blacksquare$$

Théorème 1.5.3 (*Unicité*)

Soit $U \subset \mathbb{R}^n$ un ensemble fermé et convexe et $J : U \rightarrow \mathbb{R}$ une fonction strictement convexe.

Alors il existe au plus un point de minimum de J sur U

Preuve. On va raisonner par l'absurde

Soit $x_1, x_2 \in U$, avec $x_1 \neq x_2$ deux points de minimum de J sur U . Nous avons donc :

$$J(x_1) = J(x_2) \leq J(x) \quad \forall x \in U \tag{1.2}$$

Comme J est strictement convexe, on a :

$$J\left(\frac{1}{2}x_1 + \frac{1}{2}x_2\right) < \frac{1}{2}J(x_1) + \frac{1}{2}J(x_2) = J(x_1)$$

(car $J(x_1) = J(x_2)$) et ceci est une contradiction de (1.2) ■

1.5.1 Remarques sur l'existence et l'unicité

La question de l'existence et de l'unicité d'un minimiseur x^* pour le problème (1.1) est délicate.

On peut retenir comme principe général que la compacité fournit des résultats d'existence, et la convexité un cadre favorable pour l'unicité.

Théorème 1.5.4 *Si J est continue et U est compact, alors le problème (1.1) admet au moins une solution*

Preuve. On considère (x_n) une suite minimisante, i.e. $x_n \in U$ et $J(x_n)$ converge vers la borne inférieure de J sur U $\inf_U J$ (éventuellement égale à $-\infty$). Par compacité, on peut supposer que (x_n) est convergente, quitte à extraire; notons x^* la limite. Par continuité de J , il vient que $\inf_U J = J(x^*) = \min_U J$. ■

Théorème 1.5.5 *Si J est continue et coercive, i.e. infinie à l'infini :*

$$\lim_{\|x\| \rightarrow \infty} J(x) = +\infty.$$

alors le problème (1.1) admet au moins une solution

Preuve. Soit $x_0 \in U$ et $m = J(x_0)$. Par coercivité de J , il existe une boule B centrée en x_0 en dehors de laquelle J est strictement supérieure à m . Ainsi les bornes inférieures de J sur U et sur $U \cap B$ sont les mêmes. Le résultat précédent appliqué à $U \cap B$, compact, permet de conclure ■

Théorème 1.5.6 *Si U est convexe et J strictement convexe, alors le problème (1.1) admet au plus une solution*

Preuve. Supposons que x_1^* et x_2^* soient deux solutions distinctes, alors $x = (x_1^* + x_2^*)/2 \in U$ et, par stricte convexité de J

$$J(x^*) \leq \frac{J(x_1^*) + J(x_2^*)}{2} = \min_U J.$$

qui est fournit la contradiction. ■

Chapitre 2

Quelques méthodes d'optimisation non différentiable

2.1 Introduction

Ce chapitre est consacré aux principes des méthodes les plus importantes de recherche de minimum d'une fonction convexe dans le cas de l'optimisation non différentiable.

Ces méthodes sont : la méthode de sous-gradient, la méthode de sous-gradient projeté, et les méthodes de faisceaux.

2.2 Sous-gradient et sous différentielle

Définition 2.2.1 Soit une fonction convexe $\phi : U \rightarrow \mathbb{R}$ où $U \in \mathbb{R}^n$ et un point $x^* \in U$. Un vecteur γ est un sous-gradient de ϕ en x^* si pour tout $x \in U$

$$\phi(x) \geq \phi(x^*) + \langle \gamma, x - x^* \rangle .$$

L'ensemble de tous les sous-gradients en x^* est appelé sous-différentiel de ϕ en x^* et noté

$$\partial\phi(x^*) = \{ \gamma \in \mathbb{R}^n, \forall x \in \mathbb{R}^n, \phi(x) \geq \phi(x^*) + \langle \gamma, x - x^* \rangle \}.$$

$$\mathbf{1} \quad J(x) = |x|, \partial J(x) = \begin{cases} [-1] & \text{si } x < 0 \\ [-1, 1] & \text{si } x = 0 \\ [1] & \text{si } x > 0 \end{cases}$$

2 Soit $J(x) = \|x\|$, alors pour tout g vérifiant $\|g\|_\infty \leq 1$

on a $g \in \partial J(0)$

où $\|\cdot\|_\infty$ est la norme duale de $\|\cdot\|$ (i.e, $\|x\|_\infty = \sup_{\|z\| \leq 1} \langle z, x \rangle$)

3 $J(x) = \max \{J_1(x), J_2(x)\}$ où J_1 et J_2 sont différentiable

$$\partial J(x) = \begin{cases} \{J'_1(x)\} & \text{si } J'_1(x) > J'_2(x) \\ [J'_1(x), J'_2(x)] & \text{si } J'_1(x) = J'_2(x) \\ J'_2(x) & \text{si } J'_1(x) < J'_2(x) \end{cases}$$

2.2.1 Règles de calcul de sous-différentiel

Lemme 2.2.1 Soit J une fonction convexe. On a :

i) Si J est différentiable sur son domaine, alors :

$$\forall x \in \text{int}(\text{dom}(J)), \partial J(x) = \{\nabla J(x)\} .$$

ii) Si $\text{dom}(J) \subseteq \mathbb{R}^n$, $A : \mathbb{R}^m \rightarrow \mathbb{R}^n$ un opérateur linéaire et $b \in \mathbb{R}^n$, alors la fonction $\phi(x) = J(Ax + b)$ est convexe et $\forall x \in \text{int}(\text{dom}(\phi))$, $\partial\phi(x) = A^T \partial J(Ax + b)$.

iii) Si $J(x) = \alpha_1 J_1(x) + \alpha_2 J_2(x)$ avec J_1 et J_2 convexes sur \mathbb{R}^n et $(\alpha_1, \alpha_2) \in \mathbb{R}_+ \times \mathbb{R}_+$. Alors

$$\begin{aligned} \partial J(x) &= \alpha_1 \partial J_1(x) + \alpha_2 \partial J_2(x). \\ &= \{\gamma \in \mathbb{R}^n, \exists (x_1, x_2) \in \partial J_1(x_1) \times \partial J_2(x_2), \gamma = x_1 + x_2\}. \end{aligned}$$

iv) Soit $J : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ une fonction convexe fermée et $F : \mathbb{R} \cup \{+\infty\} \rightarrow \mathbb{R} \cup \{+\infty\}$ une fonction convexe croissante. Notons : $h = F \circ J$.

Alors $\forall x \in \text{int}(\text{dom}(J))$,

$$\partial h(x) = \{\gamma_1 \gamma_2, \gamma_1 \in \partial F(J(x)), \gamma_2 \in \partial J(x)\}.$$

2.3 Méthode de sous-gradient

Les méthodes de *sous-gradient* sont les plus simples pour résoudre les problèmes d'*optimisation non différentiable*, elle sont proposées pour maximiser des fonctions concaves (ou minimiser des fonctions convexes) non nécessairement différentiables pour lesquelles il est relativement aisé de déterminer un sous-gradient en un point

Considérons le problème

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

où $U \subset \mathbb{R}^n$ et $J : U \rightarrow \mathbb{R}$ est une fonction convexe mais pas nécessairement de classe C^1

2.3.1 Direction de descente

Les algorithmes utilisés pour résoudre des problèmes d'optimisation non-différentiable consistent schématiquement à la recherche d'une direction de descente d_l et d'une longueur de pas $\tau_l > 0$. Le nouvel itéré s'obtient ensuite par $x_l = x_l + \tau_l d_l$.

2.3.2 Condition d'optimalité

Soit le problème suivant :

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

La condition d'optimalité pour un point x^* s'écrit à l'aide de l'une des trois propriétés équivalente

$$\left\{ \begin{array}{l} (i) J(x) \geq J(x^*) \text{ pour tout } x \in \mathbb{R}^n \\ (ii) 0 \in \partial J(x) \\ (iii) J'(x, d) \geq 0 \text{ pour tout } d \in \mathbb{R}^n \end{array} \right.$$

2.3.3 Algorithme de sous-gradient

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 $J(x)$ de la fonction objectif J en x et au moins un sous-gradient $\gamma \in \partial J(x)$.

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

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

où $U \subset \mathbb{R}^n$ est un ensemble convexe, $J : U \rightarrow \mathbb{R}$ est convexe et J est Lipschitz de constante L sur U .

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

Algorithme 1 : Algorithme de sous-gradient dans le cas sans contrainte. ($U = \mathbb{R}^n$).

Input :

N : un nombre d'itérations.

$x_0 \in U$: un point de départ.

Output :

x_N : une solution approchée.

Begin

For k allant de 0 à N **do**.

 Calculer $\gamma \in \partial J(x_k)$.

 Recherche linéaire : chercher un pas $\tau_k > 0$ tel que :

$$J\left(x_k - \tau_k \frac{\gamma_k}{\|\gamma_k\|}\right) < J(x_k).$$

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

End

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 $\partial f(x) = \text{signe}(x)$ 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 (1) ne soit pas de descente, un choix adéquat des τ_k rend malgré tout la méthode convergente.

Théorème 2.3.1 *Si la suite $(\tau_k)_k$ vérifie :*

$$\lim_{k \rightarrow +\infty} \tau_k = 0, \quad \sum_{k=0}^{+\infty} \tau_k = +\infty.$$

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

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

Théorème 2.3.2 *Si la suite $(\tau_k)_k$ vérifie :*

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

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

Les deux conditions $\lim_{k \rightarrow +\infty} \tau_k = 0$ et $\sum_{k=0}^{+\infty} \tau_k = +\infty$ indiquent que les pas doivent tendre vers 0, mais pas trop rapidement. Si la suite $(\tau_k)_{k \in \mathbb{N}}$ 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 $f(x) = |x|$.

2.3.4 Méthodes de sous-gradient projeté

Revenons maintenant au problème :

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

On note $P_U(x)$ la projection orthogonale du point $x \in \mathbb{R}^n$ sur le domaine admissible U . Comme l'ensemble U 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.

Algorithme 2 : Algorithme de sous-gradient dans le cas sous contrainte ($U \neq \mathbb{R}^n$).

Input :

N : un nombre d'itérations.

$x_0 \in U$: un point de départ.

Output :

x_N : une solution approchée.

Begin

For k allant de 0 à N **do**

 Calculer $\gamma \in \partial J(x_k)$.

Recherche linéaire : chercher un pas $\tau_k > 0$ tel que :

$$J\left(x_k - \tau_k \frac{\gamma_k}{\|\gamma_k\|}\right) < J(x_k).$$

$$x_{k+1} = P_U\left(x_k - \tau_k \frac{\gamma_k}{\|\gamma_k\|}\right).$$

End

Dans certains cas, nous pouvons exprimer la mise à jour du sous-programme projeté d'une autre manière.

Lorsque U est affine, c'est-à-dire que $U = \{x/Ax = b\}$, où A est grande et en plein rang, l'opérateur de projection est affine, et donné par

$$P(z) = z - A^T(AA^T)^{-1}(Az - b).$$

Dans ce cas, nous pouvons simplifier le sous-gradient par :

$$x^{(k+1)} = x^{(k)} - \tau_k (I - A^T(AA^T)^{-1}A)\gamma^{(k)}$$

2.3.5 Quelques résultats de la convergence

On donne ici quelques résultats dus la convergence de la méthode de sous-gradient.

Soit x^* une solution optimale de problème. On suppose que les sous-gradients sont bornés en norme. i.e il existe un G telle que $\|\gamma^{(k)}\| \leq G$ pour tout k . Ceci est vérifié dans le cas où J est lipschitz ($G = L$).

On suppose aussi qu'on connaît un nombre R qui satisfait $R \geq \|x^{(1)} - x^*\|$

Le nombre R peut être égale à la distance entre le point initial et l'ensemble des solutions optimales.

– On a :

$$J_{best}^{(k)} - J^* \leq \frac{R^2 + G^2 \sum_{i=0}^k \alpha_i^2}{2 \sum_{i=0}^k \alpha_i}, \text{ où } \alpha_i = \frac{\tau_i}{\|\gamma_k\|}. \quad (2.2)$$

– Si la longueur de pas est stable i.e, $\alpha_i = \frac{\tau_i}{\|\gamma_k\|}$, on obtient

$$J_{best}^{(k)} - J^* \leq \frac{R^2 + \tau^2 k}{2\tau k/G}.$$

– Le choix de séquence $\alpha_1, \dots, \alpha_k$ qui minimise (2.2) qui est la borne sup de $J_{k+1}^* - J^*$ est donné par

$$\alpha_i = \frac{R}{G\sqrt{k}}, i = 1 \dots k$$

– Une borne inférieure de J est donné par

$$J^* \geq l_k = \frac{2 \sum_{i=1}^k \alpha_i J(x^{(i)}) - R^2 \sum_{i=1}^k \alpha_i^2 \|\gamma^{(i)}\|^2}{2 \sum_{i=1}^k \alpha_i}.$$

peut être utilisé comme critère d'arrêt de l'algorithme.

– L'algorithme s'arrête quand $J_{best}^{(k)} - l_{best}^{(k)}$ est le plus petit qu'un seuil. où

$$l_{best}^{(k)} = \max \{l_1, \dots, l_k\}$$

Théorème 2.3.3 On note $J_k^* = \min_{k \in \{0, \dots, k\}} J(x_k)$ et R le diamètre de l'ensemble U . Alors :

$$J_k^* - J^* \leq L \frac{R^2 + \sum_{i=0}^k \tau_i^2}{2 \sum_{i=0}^k \tau_i}$$

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

$$J_k^* - J^* \leq \frac{LR}{\sqrt{N+1}}$$

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

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

or par définition du sous-différentiel :

$$J(y) \geq J(x) + \langle \gamma, y - x \rangle$$

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

$$|\langle \gamma, y - x \rangle| \leq L \|x - y\|, \forall y \in U$$

En particulier pour $y = \gamma + x$ on obtient $\|\gamma\| \leq L$. 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 :

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|x_k - x^* - \tau_k \frac{\gamma_k}{\|\gamma_k\|}\|^2 \\ &= \|x_k - x^*\|^2 + \tau_k^2 - 2\tau_k \left\langle x_k - x^*, \frac{\gamma_k}{\|\gamma_k\|} \right\rangle \end{aligned}$$

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

$$\|x_{k+1} - x^*\|^2 \leq \|x_k - x^*\|^2 + \tau_k^2 - 2 \frac{\tau_k}{\|\gamma_k\|} (J(x_k) - J(x^*))$$

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

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

D'où :

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

Ce qui permet de conclure :

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

■

2.3.6 Choix de polyak

Polyak ([8]) à proposé un choix pour la longueur de pas $\alpha_k = \frac{\tau_k}{\|\gamma_k\|^2}$ dans les deux cas suivant :

– Cas où J^* est connue : il a proposé

$$\alpha_k = \frac{J(x^k) - J^*}{\|\gamma^{(k)}\|^2}.$$

– Cas où J^* est inconnue :

$$\alpha_k = \frac{J(x^k) - J_{best}^k + \tau_k}{\|\gamma^{(k)}\|^2}.$$

2.4 Méthodes des plans sécants et de faisceaux

Les méthodes de faisceaux sont considérées parmi les plus efficaces pour la résolution de problème non-différentiables. Historiquement, les méthodes de faisceaux sont basées sur la méthode des plans sécants voir ([7]) 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 J pour construire un modèle (convexe) de J , plus facile à minimiser.

Etant donné un faisceau d'informations :

$$\forall y \in \mathbb{R}^n, \phi_k(y) = \max_{1 \leq i \leq k} \{J_i + \langle \gamma_i, x - x_i \rangle\} (\leq J(y))$$

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.

Idée des plans sécants : Accumuler des informations des itérations précédentes : J

$$J_i = J(x_i) \text{ et } \gamma_i \in \partial J(x_i), 1 \leq i \leq k$$

A l'itération k on considère \tilde{J} une approximation affine par morceaux de J (plus facile à minimiser) :

$$\phi_k(y) = \max_{1 \leq i \leq k} \{J_i + \langle \gamma_i, x - x_i \rangle\}$$

Notons que $\phi_k \leq \phi_{k+1}$ et $\phi_k(x) \leq J(x), \forall k \geq 0, \forall x \in \mathbb{R}^n$. On détermine $C \in \mathbb{R}^n$ compact

contenant un minimiseur de J et l'on fixe $\varepsilon > 0$ critère d'arrêt. Pour $k > 0$

$$\begin{aligned} \varepsilon_k &= J(x_k) - \phi(x_k); && \text{arrêter si } \varepsilon_k \leq \varepsilon \\ z_{k+1} &= \arg \min_{x \in C} \phi(x) \\ d_k &= -z_{k+1} + x_k && \text{(direction de recherche)} \\ x_{k+1} &= x_k - \tau_k d_k && (\tau_k > 0 \text{ obtenu par recherche linéaire}). \end{aligned}$$

d_k n'est pas forcément une direction de descente pour J .

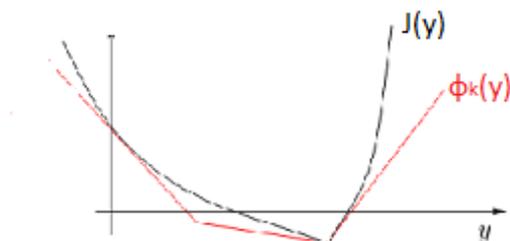


FIG. 2.1 – Méthode de plans sécants -construction d'un modèle convexe linéaire par morceaux d'une fonction convexe non différentiable

2.5 Exemples

2.5.1 Méthode de sous-gradient

Nous considérons le problème de la minimisation d'une fonction linéaire par morceaux

$$\text{minimiser } f(x) = \max_{i=1, \dots, m} (a_i^T x + b_i)$$

avec la variable $x \in \mathbb{R}^n$. Bien sûr, ce problème est facilement (et efficacement) peut être

résolu par une programme linéaire.

Trouver un sous-gradient de f est facile : étant donné x , on trouve d'abord un indice j pour lequel

$$a_j^T x + b_j = \max_{i=1, \dots, m} (a_i^T x + b_i).$$

Ensuite, nous pouvons prendre comme un sous-gradient $\gamma = a_j$. On peut prendre

$$G = \max_{i=1, \dots, m} \|a_i\|.$$

Nous illustrons la méthode du sous-gradient pour $n = 20$ variables et $m = 100$ termes, avec les données a_i et b_i qui nous commençons par $x^{(1)} = 0$. Il n'y a pas de moyen simple de trouver une valeur justifiable pour R (i.e, une valeur de R pour laquelle on peut prouver que $\|x^{(1)} - x^*\|_2 \leq R$), Nous prenons donc $R = 10$. Pour notre cas il se trouve que $\|x^{(1)} - x^*\| = 0.91$, où nous avons calculé la valeur optimale $f^* \approx 1.1$ en utilisant la programmation linéaire.

Considérons d'abord la règle de $\alpha_k = \frac{\tau}{\|\gamma_k\|_2}$. La figure (2.2) montre la convergence de $f_{best}^{(k)} - f^*$ pour $\tau = 0.05, \tau = 0.01, \tau = 0.005$. La figure révèle un compromis : τ plus grand donne une convergence plus rapide, mais une optimalité finale plus grande.

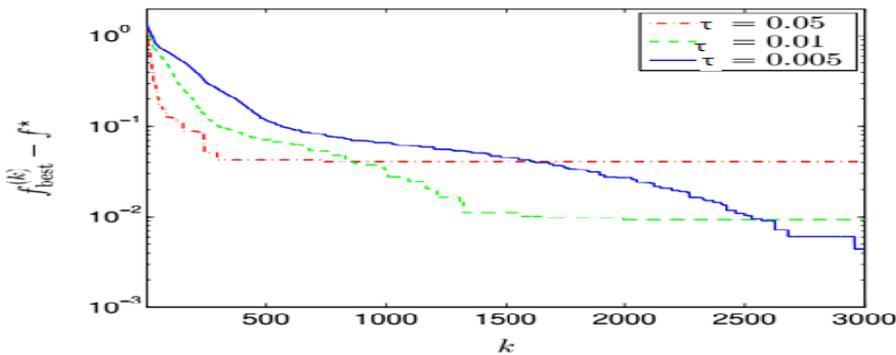


FIG. 2.2 – la valeur de $f_{best} - f^*$ en fonction de nombre d'itérations k par la méthode du sous-gradient avec une longueur de pas constante τ

La figure (2.3) montre la progression de la méthode de sous-gradient par le choix de Polyak sans connaître la valeur de f^* .

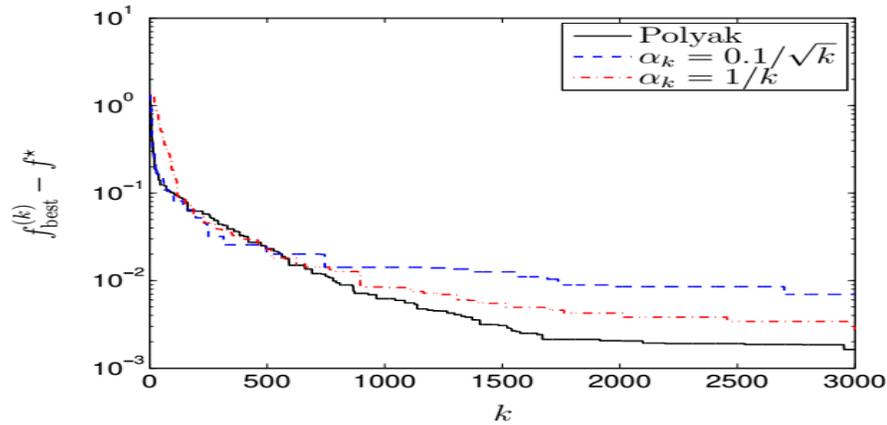


FIG. 2.3 – La valeur de $f_{best}^{(k)} - f^*$ en fonction du nombre d'itérations k , pour le sous-gradient avec le choix de Polyak (ligne noire continue) et les méthodes de sous-gradient avec une taille de pas décroissante.

2.5.2 Méthode sous-gradient projeté

Nous considérons le problème

$$\begin{aligned} & \text{minimiser} && \|x\|_1 \\ & \text{sous réserve} && Ax = b \end{aligned}$$

Où la variable $x \in \mathbb{R}^n$ et les données $A \in \mathbb{R}^{n \times m}$ et $b \in \mathbb{R}^m$

Nous supposons que $\text{rang}(A) = m$ tel que $m < n$. Bien sûr, ce problème est facilement résolu en utilisant la programmation linéaire.

Un sous-gradient de l'objectif en x est donné par $\gamma = \text{signe}(x)$. Ainsi, le sous-gradient projetée :

$$x^{(k+1)} = x^{(k)} - \alpha_k (I - A^T(AA^T)^{-1}A) \text{signe}(x^{(k)})$$

Pour $n = 1000$ et $m = 50$.Nous commençons par le point de départ :

$$x^{(1)} = A^T(AA^T)^{-1}b.$$

Nous pouvons résoudre ce problème par la programmation linéaire, on trouve $f^* \approx 3.2$

.La figure (2.4) montre l'avancement du méthode de sous-gradient projeté avec le choix de

Polyak $\tau_k = \frac{100}{k}$.

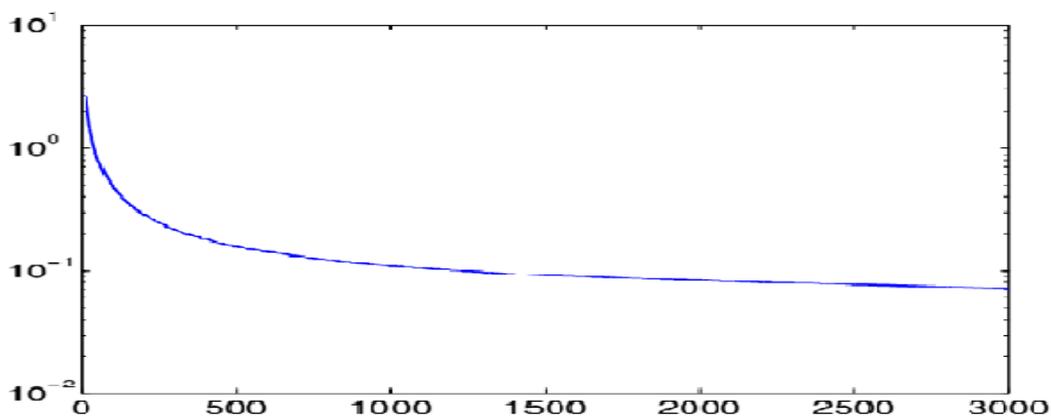


FIG. 2.4 – la valeur de $f_{best}^{(k)} - f^*$ par rapport au nombre d'itérations k , pour la méthode du sous-gradient avec le choix de Polyak $\tau_k = \frac{100}{k}$.

Bibliographie

- [1] Bruno Galerne, *Optimisation (MML1E31) Notes de cours*, Master 1 Mathématiques et Modélisation (MM), 2017-2018.
- [2] D. Verwaerde et P. Laurent-Gengoux, *Mathématiques 2, Optimisation*, Ecole Centrale Paris, 2006-2007.
- [3] Guillaume Laurent, *Optimisation sans contrainte de fonctions continues non linéaires*, 2012.
- [4] Ionel Sorin CIUPERCA, *Cours optimisation*, Cours à l'ISFA, en M1SAF.
- [5] Aude Rondepierre, *Introduction à l'optimisation convexe non différentiable*.
- [6] Stephen Boyd (with help from Jaehyun Park), *Subgradient Methods* Notes for EE364b, Stanford University, Spring 2013–14 May 2014; based on notes from January 2007.
- [7] 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.
- [8] .B. Polyak, *Introduction to Optimization*, Optimization Software, Inc, 1987.

Annexe : Abréviations et Notations

Les différentes abréviations et notations utilisées tout au long de ce mémoire sont expliquées ci-dessous :

C^1 : L'ensemble des fonction différentiable et les dirivées premiers continue.

J_d : La dirivée directionnel dans la direction d .

$\| \cdot \|$: Désigne la norme eucludienne dans \mathbb{R}^n .

$\nabla J(x)$: Le gradient de la fonction J .en x .

$\partial J(x)$: Sous-différentiel de la fonction J .en x .

γ : Le sous-gradient

P_U : Opérateur de projection sur l'ensemble U ..