SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Convex Optimization and Applications

By

Stella-Varvara C. Gkila

A THESIS

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfilment of the requirements for

the degree of Master of Science in Statistics

Athens, Greece
June 2019


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

Κυρτή Βελτιστοποίηση και Εφαρμογές

Στέλλα-Βαρβάρα Χ. Γκίλα

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Ιούνιος 2019


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

DEDICATION

To my family.

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor, Athanasioss Yannacopoulos who was very supportive and with his guidance helped me to conduct my thesis. Additionally, all my teachers over the years and my family and friends who have always been here for me.

VITA

My name is Stella-Varvara Gkila and I born in Athens, Greece. I studied maths. I'm interested in maths and their applications and I will be a M.Sc. Statistics graduate.

ABSTRACT

Stella-Varvara Gkila

Convex Optimization and Applications.

June 2019

In the following thesis, we discuss algorithms for convex optimization. Is optimization for convex function on convex sets. These algorithms are based on notion of functional and convex analysis. We use functional analysis to construct sequence which are convergent in Hilbert space and $\mathbb{R}^n$. The basic idea is that the iterative sequence we construct converges to the minimum of objective function. We generalize the notion of gradient and differentiable functions for non-smooth, so we can minimize them. The first method we see is the gradient method, which is about convex and differentiable functions. Next algorithm, proximal point is about non-smooth functions and then we combine gradient and proximal and we have an algorithm for functions, which is the sum of smooth and non-smooth. Finally, we study the primal dual algorithm. An example of these methods is provided to Lasso function.

ΠΕΡΙΛΗΨΗ

Στέλλα-Βαρβάρα Γκίλα

Κυρτή Βελτιστοποίηση και εφαρμογές.

Ιούνιος 2019

Στην παρούσα εργασία, θα παρουσιάσουμε αλγορίθμους για την ελαχιστοποίηση κυρτών συναρτήσεων πάνω σε κυρτά σύνολα. Αυτοί οι αλγόριθμοι βασίζονται σε έννοιες της συναρτησιακής και κυρτής ανάλυσης. Χρησιμοποιούμε βασικές έννοιες και θεωρήματα της ανάλυσης για πετύχουμε την σύγκλιση των ακολουθιών που κατασκευάστηκαν. Η βασική ιδέα είναι να φτιάξουμε επαναληπτικές διαδικασίες που συγκλίνουν στο ελάχιστο της αντικειμενικής συνάρτησης. Ο χώρος που θα δουλέψουμε περισσότερο είναι ο Hilbert με κάποια παραδείγματα και αναφορές στον $\mathbb{R}^n$. Θα γενικεύσουμε την έννοια της παραγώγου με σκοπό να μπορούμε να διαχειριστούμε συναρτήσεις κυρτές αλλά όχι διαφορίσιμες. Η πρώτη μέθοδος είναι η gradient method, αφορά παραγωγίσιμες συναρτήσεις. Μετά θα αναφερθούμε στον proximal point που αφορά μη παραγωγίσιμες συναρτήσεις. Οι δύο παραπάνω μέθοδοι συνδυάζονται και μας δίνουν τον proximal gradient method, με τον οποίο μπορούμε να ελαχιστοποιήσουμε συναρτήσεις που εμπλέκονται όροι διαφρορίσιμων συναρτήσεων και μη. Μετά θα αναφερθούμε στον primal-dual αλγόριθμο. Τέλος θα παρουσιάσουμε ένα παράδειγμα των παραπάνω αλγορίθμων στο πρόβλημα ελαχιστοποίησης Lasso.

Table of Contents

1 INTRODUCTION ..... - 1 -

2 ANALYSIS ..... - 3 -

2.1. Norms ..... *- 3 -*

2.2. Sequences ..... *- 5 -*

2.3. Topological Properties ..... *- 6 -*

2.4. Functions ..... *- 7 -*

3 CONVEXITY ..... - 11 -

3.1. Convex Sets ..... *- 11 -*

3.2. Affine Sets ..... *- 12 -*

3.3. Separating Theorems ..... *- 13 -*

3.4 Convexity and Nonexpansiveness. ..... *- 13 -*

4 CONVEX ANALYSIS AND SUBDIFFERENTIAL CALCULUS ..... - 17 -

4.1. Convex Function ..... *- 17 -*

4.2. Convexity and continuity ..... *- 19 -*

4.3. Convexity and Differentiability. ..... *- 19 -*

4.4. Subgradients ..... *- 23 -*

4.5. Subdifferential Calculus ..... *- 28 -*

4.6. Proximal Map and Moreau - Yosida Regularization. ..... *- 30 -*

4.7. The Legendre – Fenchel conjugate ..... *- 33 -*

4.8. Fenchel – Rockafellar Duality ..... *- 35 -*

5 ALGORITHMS ..... - 37 -

5.1. Iterative Procedures ..... *- 37 -*

5.2. Gradient Method ..... *- 37 -*

5.3. Proximal Point Algorithm ..... *- 41 -*

5.4. Proximal Gradient Method ..... *- 45 -*

5.5. Accelerated proximal gradient. ..... *- 47 -*

5.6. Primal dual Algorithm. ..... *- 47 -*

6 Minimization of Lasso function ..... - 49 -

6.1. LASSO ..... *- 49 -*

6.2. Proximal gradient method ..... *- 50 -*

6.3. Primal-Dual Problem ..... *- 50 -*

APPENDICES ..... - 52 -

References ..... - 55 -

LIST OF FIGURES

Figure 1 lower-semicontinuous function. ..... - 8 -
Figure 2 Geometrical Interpretation of contraction and nonexpansiveness ..... - 14 -
Figure 3 Convex function. ..... - 17 -
Figure 4 Fermat's rule. The vector is the gradient of f ..... - 21 -
Figure 5 Geometrical Interpretation of convex differentiable function. ..... - 23 -
Figure 6 Geometrical interpretation of subgradients ..... - 24 -
Figure 7 Geometry of Conjugate ..... - 35 -
Figure 8 Gradient Method ..... - 38 -
Figure 9 A proximal point of Gradient method ..... - 40 -
Figure 10 Interpretation of proximal operator ..... - 43 -
Figure 11 Proximal Operator of Projection ..... - 43 -


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

1 INTRODUCTION

The main issue in this thesis is theory of convex optimization. We study about main notion of functional and convex analysis and their usability to optimization. In general, we need to construct sequences with good properties like monotonicity and convergence. We use these sequences to minimize convex functions which are differentiable or not. So, we generalize the notion of derivative of a function.

In chapter 2 we define the basic notion of functional analysis, like norms, convergence sequence, Banach and Hilbert spaces. We refer to basic properties of Hilbert spaces, some of them are weakly convergence, projection. We define the function extended real set and we define the domain, sublevel sets, graph and epigraph of function. We connect lower-semicontinuity with the epigraph and sublevel sets. In the end we study about minimizing sequence of a function and how convex functions minimizing in a reflexive space (i.e. Hilbert space).

In chapter 3, we study the convexity of sets. We define affine sets hyperplanes and half-spaces and finally, the Hanh – Banach Theorem.

In chapter 4, we define convex function and their connection with convex sets. Then we connect convexity with differentiability, and we generalize all the properties convex differentiable functions for nondifferentiable functions. The generalization of gradient is the subgradient and is the notion of subgradient. Next, we introduce the Moreau – Yosida regularization, a function which is a smooth version of non-smooth function. We know from previous theory that this function has unique minimizer in Hilbert spaces and this unique minimizer of Moreau – Yosida regularization is called proximal operator. Proximal operator has very good properties, is monotone operator, is nonexpansive and we can interpret it as the resolvent of subgradient. Then we study the Fenchel conjugate and how we use it in duality. Finally, we define the dual problem.

In chapter 5, we will analyze the idea of iterative algorithms. First, we study gradient method, which is about differentiable functions. This method exploits the monotonicity of the gradient and construct a sequence which in Hilbert space converges to the minimum of, f if exists. Next, we study the proximal point method, which is a generalization of gradient for non-smooth functions. Then we combine the two previous methods and we have the proximal gradient method, which is about functions, where

involve smooth and non-smooth functions. The basic idea in all three algorithms is to construct sequences, where are converge under assumptions to the minimum of f. The last algorithm we present is the primal dual, which is for smooth and non-smooth functions and uses the conjugate theory.

In the chapter 6, we apply proximal gradient method on Lasso function. This thesis aims to address the theory of convex optimization presenting the main points of the works of [1] Juan Peypouquet (Convex Optimization in Normed Spaces,2015), [2] Heinz H. Bauschke, Patrick L. Combettes (Convex Analysis and Monotone operator Theory in Hilbert Spaces, 2010), Stephen Boyd, Lieven Vandeberghe and R. Tyrrell Rockafellar.

2 ANALYSIS

2. Introduction

General convex analysis and functional analysis are very close related. In this section we discuss basic notions of functional analysis like norms, normed spaces, inner product, Banach spaces and Hilbert spaces, basic convergence of sequence, topological properties. Basic notions of functions like epigraph of functions. We define the extended real line, and proper and lower-semicontinuous functions. Also, we study the minimizing of functions in reflexive spaces.

2.1. Norms

Definition 2.1.1. Let $A \subseteq \mathbb{R}^N$ a real vector space. Each function $\|\,.\,\|: A \to \mathbb{R}$ with the following properties is a *norm* on A :

(a) $\|x\| \geq 0$ for each $x \in A$ and $\|x\| = 0$ if and only if $x = 0$.

(b) $\|\lambda x\| = |\lambda|\|x\|$ for each $\lambda \in \mathbb{R}$ and each $x \in A$.

(c) $\|x + y\| \leq \|x\| + \|y\|$ for each $x, y \in$ A. (*triangle inequality*).

If function $\|\,.\,\|$ is a norm on X, the pair $(X, \|\,.\,\|)$ is called *normed space.* ⊡

Note that a norm is a measure of the length of a vector and a distance between two vectors.

Spaces with finite dimensions.

1. We define on $\mathbb{R}^m$ the *supremum norm* $\|\,.\,\|_\infty : \mathbb{R}^m \to \mathbb{R}$ as:

$$\|x\|_\infty := max\,\{|x_i|: i = 1, \dots, m\}$$

The space $(\mathbb{R}^m, \|\,.\,\|_\infty)$ is denoted ℓ_∞^m.

2. We define on $\mathbb{R}^m$ the $\ell_1 - norm$ $\|.\|_1 : \mathbb{R}^m \to \mathbb{R}$ as:

$$\|x\|_1 = |x_1| + \cdots + |x_m|$$

The space $(\mathbb{R}^m, \|\,.\,\|_1)$ is denoted ℓ_1^m.

3. We define on $\mathbb{R}^m$ the *Euclidean norm* $\|\,.\,\|_2 : \mathbb{R}^m \to \mathbb{R}$ like:

$$\|x\|_2 := \left(\sum_{i=1}^{m} |x_i|^2\right)^{1/2}$$

Proposition 2.1.2. (Cauchy – Schwarz inequality). Let $x, y \in \mathbb{R}^m$, then we have,

$$\sum_{i=1}^{m} |x_i y_i| \leq \left(\sum_{i=1}^{m} |x_i|^2 \right)^{1/2} \left(\sum_{i=1}^{m} |y_i|^2 \right)^{1/2}$$

Proof. If we set $A = \sum_{i=1}^{m} |x_i y_i|$, $B = (\sum_{i=1}^{m} |x_i|^2), C = (\sum_{i=1}^{m} |y_i|^2)$. We have to prove that $A^2 \leq BC \Leftrightarrow (2A)^2 \leq 4BC \Leftrightarrow (2A)^2 - 4BC \leq 0$ .We suppose the function $g: \mathbb{R} \to \mathbb{R}$ $g(\lambda) := (\lambda|x_1| + |y_1|)^2 + \cdots + (\lambda|x_m| + |y_m|)^2 \geq 0$, which after operations, takes the following form $g(\lambda) = B\lambda^2 + 2A\lambda + C \geq 0$, for each $\lambda \in \mathbb{R}$. If $A = 0$, then $x_i = 0$ for each $i = 1, \dots, m$ and the inequality holds (as equality). After all we suppose that $A > 0$ and then $g(\lambda) > 0$ for each $\lambda \in \mathbb{R}$ and the quantity $D = (2A)^2 - 4BC \geq 0$ and we have the inequality. ⊡

Definition 2.1.3. An *inner product* on X (linear vector space) is a function $\langle . , . \rangle : X \times X \to \mathbb{R}$ such that:

(a) $\langle x, x \rangle \geq 0 \forall x \in \mathbb{R}$

(b) $\langle x, x \rangle = 0 \Leftrightarrow x = 0$

(c) $\langle x, y \rangle = \langle y, x \rangle$

(d) $\langle x, \lambda_1 y_1 + \lambda_2 y_2 \rangle = \lambda_1 \langle x, y_1 \rangle + \lambda_2 \langle x, y_2 \rangle \; \forall x, y \in \mathbb{R}$. ⊡

The most common example is $\langle x, y \rangle = \sum_{j=1}^{m} x_j y_j$. We observe that $\langle x, x \rangle = \|x\|_2{}^2 \forall x \in \mathbb{R}$.

Definition 2.1.4. The *dual norm* of $\| . \|$ is denoted $\| . \|_*$ and is defined as

$$\|y\|_* = sup \{ \langle y, x \rangle | \|x\| \leq 1 \}. ⊡$$

Examples:

1. The dual of the dual norm is the original norm, $\|x\|_{**} = \|x\|$ for all x.
2. The dual of Euclidean norm is the Euclidean norm.
3. The dual of $\ell_1 -$ norm is the $\ell_\infty -$ norm, and the opposite. Since,

$$sup \{ \langle y, x \rangle \mid \|x\|_\infty \leq 1 \} = \sum_{i=1}^{n} |y_i| = \|y\|_1.$$

∘We denoted $X^* = \mathcal{J}(X, Y)$ the space of bounded linear operators from space $(X, \| . \|_X)$ to $(Y, \| . \|_Y)$. A linear operator $K : X \to Y$ is *bounded* if:

$$\|K\|_{X^*} = sup_{\|x\|_X = 1} \|K(x)\|_Y < \infty.$$

∘The *topological dual* of a normed space $(X, \| . \|)$ is the normed space $(X^*, \| . \|_*)$, where $\| . \|_* = \| . \|_{\mathcal{J}(X;Y)}$.

∘The function $\langle .,. \rangle_{X^*,X} : X^* \times X \to \mathbb{R}$, defined as $\langle K, x \rangle_{X^*,X} = K(x)$ is called *bilinear* function, and is the duality product between X and $X*$.

∘The topological dual of $(X^*, \|.\|_*)$, is denoted $(X^{**}, \|.\|_{**})$ and is called the *topological bidual* of $(X^*, \|.\|_*)$. We define the function $\mu: X \to \mathbb{R}$ as $\mu_x(K) = \langle K, x \rangle_{X^*,X}$, $\forall K \in X^*$.

Definition 2.1.5. We called the *canonical embedding* of X into X^{**} the function $\mathcal{J}: X \to X^{**}$, defined by $\mathcal{J}(x) = \mu_x$. ⊡

Definition 2.1.6. A normed space $(X, \| \, . \|)$ is *reflexive* if for the canonical embedding we have, $\mathcal{J}(X) = X^{**}$. ⊡

2.2. Sequences

Definition 2.2.1. A *sequence* is a function $x: \mathbb{N} \to \mathbb{R}$. We denote $x_n := x(n)$ or $\{x_n\}_{n=1}^{\infty}$.

Definition 2.2.2. Let a normed space $(X, \| \, . \|)$. A sequence x_n in X *(strongly) converges* to $\overline{x} \in X$, and we write $x_n \to \overline{x}$ as $n \to \infty$ if $\lim_{n\to\infty} \|x_n - \overline{x}\| = 0$. We say that the *limit* of the sequence x_n is $\overline{x}$. ⊡

Definition 2.2.3. Let a normed space $(X, \| \, . \|)$. A sequence $x_n$is called *Cauchy sequence* if for each m, n we have $\lim_{m,n\to\infty} \|x_m - x_n\| = 0$. ⊡

Proposition 2.2.4. Let x_n a convergent sequence, then the sequence is Cauchy. ⊡

Proposition 2.2.5. Every Cauchy is sequence is bounded. ⊡

Proposition 2.2.6. Let $(X, \| \, . \|)$ a normed space. If every Cauchy sequence is convergent the normed space we say that is complete and the normed space is called *Banach space*. ⊡

Proposition 2.2.7. Let X space with inner product. The function $\| \, . \|: X \to \mathbb{R}$, where $\|x\| = \sqrt{\langle x, x \rangle}$ is a norm.

Definition 2.2.8. A real vector space X with inner product is called *Hilbert space* $\mathbb{H}$ if X is complete to norm $\| \, . \|$ which is associated with inner product.

Examples 2.2.9. 1) Every Hilbert space from definition above is Banach space.

2) The Euclidean space $\mathbb{R}^n$ is a Hilbert space, with the norm $\|x\| = \sqrt{\sum_k x_k^2}$.

Weakly Convergent Sequences

Definition 2.2.10. Let a normed space $(X, \| . \|)$. A sequence x_n in X *converges weakly* to $\overline{x}$, as $n \to \infty$ if for each $f \in X^*$, we have

$$lim_{n\to\infty} f(x_n) = f(\bar{x}).$$

In this case, the weakly convergence of x_n means, convergence of $f(x_n)$ to $f(\bar{x})$ for each $f \in X^*$.

Note that a convergent sequence is converges weakly, since

$$|\langle f, x_n - \overline{x}\rangle| \leq \|f\|_* \|x_n - \overline{x}\|.$$

Lemma 2.2.11. For each $\alpha \in \mathbb{H}$, the function $f_\alpha: \mathbb{H} \to \mathbb{R}$ with $f_\alpha(x) = \langle x, \alpha\rangle \in \mathbb{H}^*$ and $\|f_\alpha\|_{\mathbb{H}^*} = \|\alpha\|_{\mathbb{H}}$.

Theorem 2.2.12. (Riesz Representation Theorem) Let $\mathbb{H}$ Hilbert space, and $f \in \mathbb{H}^*$. Then, there are unique $a \in \mathbb{H}$ such that $f = f_\alpha$.

Proposition 2.2.13. Let $\mathbb{H}$ is a Hilbert space. Then a sequence $(x_n) \in \mathbb{H}$ converges to $\hat{x}$, if and only if, $\langle x_n, z\rangle \to \langle \hat{x}, z\rangle$.

Proof. From definition 2.2.10 and Theorem 2.2.12. we have the conclude. ⊡

Corollary 2.2.14. Hilbert spaces is reflexive.

Proof. We take $a \in \mathbb{H}^{**}$, and from Riesz Representation theorem we have $y \in \mathbb{H}^*$ such that $a_y = \langle z, y\rangle_*$, for each $z \in \mathbb{H}^*$, and then $b_z \in \mathbb{H}$ such that $y = \langle b_z, x\rangle$ for all $x \in \mathbb{H}$. Therefore, $a = \langle y, z\rangle_* = z(b_z)\ \forall z \in \mathbb{H}^*$. ⊡

An important property of Hilbert spaces is the notion of projection.

Proposition 2.2.15. Let $C \subset \mathbb{H}$, $C \neq \emptyset$ closed and convex. Let $x \in \mathbb{H}$. Then there exists a unique point $y^* \in C$ such that

$$\|x - y^*\| = min_{y\in C}\|x - y\|.$$

Additionally, it is the only element of K such that

$$\langle x - y^*, y - y^*\rangle \leq 0, \text{ for all } y \in C.$$

This property means that there is a unique point y^* in C which is closest to $x \in \mathbb{H}$ [1].

2.3. Topological Properties

Definition 2.3.1. Let $(X, \| . \|)$ be a normed space and let a point $x_0 \in X$.

(a) The *open ball* with center the point x_0 and radius $r > 0$ is the set

$$B_X(x_0, r) = \{\, x \in X: \|x - x_0\| < r\}\,.$$

(b) The *closed ball* with center the point x_0 and radius $r > 0$ is the set

$$\overline{B}_X(x_0, r) = x \in X: \|x - x_0\| \leq r. \boxdot$$

Definition 2.3.2. Let $(X, \|\,.\,\|)$ be a normed space and let A $\subseteq X$. The element $x \in A$ is called an *interior point* of A if there exists a $r > 0$ such that $B_X(x_0, r) \subseteq A$. The set of all points interior to A is called the *interior* of A and is denoted $\boldsymbol{int}\ A$.⊡

Definition 2.3.3. Let $(X, \|\,.\,\|)$ be a normed space and let A $\subseteq X$.

(a) The set A is called *open* if every element in A is an interior point.

(b) The set A is called *closed* if its complement $A^c = X \backslash A$ is open. ⊡

Definition 2.3.4. Let $(X, \|\,.\,\|)$ be a normed space and let A $\subseteq X$.

(a) The element $x \in X$ is called *contact point* of A if $\forall\ \varepsilon > 0$ it holds :

$$A \cap B_X(x, \varepsilon) \neq \emptyset.$$

(b) The *closure* of A, is the set of all contact points to A

$$cl(A) = \{x \in X: \forall \varepsilon > 0, A \cap B(x, \varepsilon) \neq \emptyset\}.⊡$$

Let (X, τ) be a topological vector space. The *weak topology* on X^* (dual) is defined to be the coarsest topology (the one with the fewest open sets) under which element $x \in X$ correspond to a continuous map on X^*.

Definition 2.3.5. The topological space (X, τ) is *Hausdorff*, if for each pair $x \neq y$, there are open and disjoint set on X, $G \cap H = \emptyset$, such that $x \in G$, $y \in H$. ⊡

2.4. Functions

The Extended Real Line

The extended real line $[-\infty, +\infty] = \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$. We join the elements $-\infty, +\infty$ to the real line $\mathbb{R}$ and we extend the order for each $\xi \in \mathbb{R}\ -\infty < \xi < +\infty$. We can define function on a set X with values only in $\mathbb{R} \cup \{+\infty\}$ or in $\mathbb{R} \cup \{-\infty\}$.

Example. The indicator function of $A \subset X$, is defined as,

$$\delta_A(x) = \begin{cases} 0\,, x \in A \\ +\infty, otherwise \end{cases}.$$

These function is very useful because we can define the optimization problem for a function $f: X \to \mathbb{R}$, $min\,\{f(x): x \in A\}$ like $min\,\{f(x) + \delta_A(x): x \in X\}$. The second problem has better properties. Like linearity.

Definition 2.4.1. A function $f(x)$ is called *Lipschitz continuous* on X if:

$$\|f(x) - f(y)\| \leq L\|x - y\|\ \forall x, y \in X.⊡$$

Definition 2.4.2. Let X be a nonempty set and let $f: X \to [-\infty, +\infty]$.

(a) The *domain* of f is $dom(f) = \{x \in X | f(x) < +\infty\}$, is the set of points where f is finite.

(b) The function f is *proper* if $dom(f) \neq \emptyset$.

(c) Given $\gamma \in \mathbb{R}$, the γ −sublevel set of f is $\Gamma_\gamma(f) = \{x \in X | f(x) \leq \gamma\}$.

(d) The *graph* of f is $graf = \{(x, \alpha) \in X \times \mathbb{R} | f(x) = \alpha\}$.

(e) The *epigraph* of f is $epif = \{(x, \alpha) \in X \times \mathbb{R} | f(x) \leq \alpha\}$.

(f) The function f is *inf-compact*, if for each $\gamma \in \mathbb{R}$ the $\Gamma_\gamma(f)$ is relatively compact. (The closure of sublevel is compact) ⊡

The epigraph includes the graph of f and all points above it.

We define $argmin(f) = \{x^* \in X : f(x^*) \leq f(x)\ forall\ x \in X.\}$ ⊡

We observe that if $x \in dom(f)$, then $x \in \Gamma_{f(x)}(f)$ and that

$$argmin(f) = \cap \Gamma_\gamma(f), \text{ for } \gamma > inf(f).$$

Let (X, τ) is a Hausdorff space. A function $f : X \to \mathbb{R} \cup \{+\infty\}$ is *lower-semicontinuous* at a point $x_0 \in X$ if for each $\alpha < f(x_0)$ there is a neighborhood V of x_0 such that $f(y) > \alpha$ for all $y \in V$. If f is lower-semicontinuous at every point of X, we say that f is *lower-semicontinuous* in X. ⊡

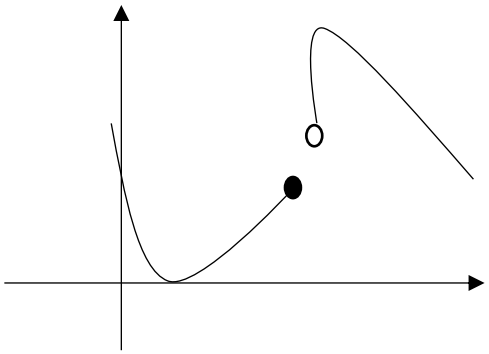


Figure 1 lower-semicontinuous function.

Let $f : X \to [-\infty, +\infty]$ and let $A \subset X$.

(a)The *infimum* of f over A is denoted inf $f(A)$ or $\inf_{x \epsilon A} f(x)$.

(b) The supremum of f over A is denoted sup $f(A)$ or $\sup_{x \epsilon A} f(x)$.

The definition 2.4.4. is equivalent with the next one.

Definition 2.4.6. Let an extended real valued function $f : X \to [-\infty, +\infty]$ it is *lower-semicontinuous* (l.s.c.) if, for all $x \in X$, if $x_n \to x$, then $f(x) \leq \lim_{n \to \infty} inf f(x_n)$. ⊡

Theorem 2.4.7. Let $f : X \to [-\infty, +\infty]$. The following statements are equivalent:

(a) The function f is l.s.c.

(b) The set $epi(f)$ is closed in $X \times \mathbb{R}$

(c) For each $\gamma \in \mathbb{R}$, the γ −sublevel set is also closed.

Proof. (a) ⇒ (b)

Let f l.s.c. and take an element $(x_0, \alpha) \notin epi(f)$. From the definition 2.4.4. of l.s.c., we have that $\alpha < f(x_0)$. We take an element $y \in (\alpha, f(x_0))$ and from l.sc. we have a neighborhood V of x_0 such that $f(z) > y$ for all $z \in V$. From all this is obvious that the set $V \times (-\infty, y)$ is a neighborhood of (x_0, α), where the intersection with $epi(f)$ is the empty. So, the set $epi(f)$ is closed.

(b) ⇒ (c)

Let $epi(f)$ is closed. For each $\gamma \in \mathbb{R}$, the $\gamma -$ sublevel set of f is homeomorphic to $epi(f) \cap [X \times \gamma]$. And from that $\Gamma_\gamma(f)$ is closed.

(c) ⇒ (a)

Let $\Gamma_\gamma(f)$ be closed and take a random $x_0 \in X$ and $\alpha \in \mathbb{R}$ such that $\alpha < f(x_0)$. Then $x_0 \notin \Gamma_\alpha(f)$ and because the sublevel set is closed, there is a neighborhood $V(x_0)$ that the intersection with $\Gamma_\alpha(f)$ is empty. So, $f(z) > \alpha$ for all $z \in V(x_0)$. ⊡

Lemma 2.4.8. Let $(f_i)_{i \in I}$ be a family of function from X to extended real line. Then we have the following statements:

(a) $epi(sup_{i \in I} f_i) = \cap_{i \in I} epi(f_i)$

(b) If I is finite, then $epi(min_{i \in I} f_i) = \cup_{i \in I} epi(f_i)$

Proof. [2] (a) Let $(x, \alpha) \in X \times \mathbb{R}$ and $(x, \alpha) \in epi(sup_{i \in I} f_i)$, which means that $sup_{i \in I} f_i(x) \leq \alpha$ so for each $i \in I$, $f_i(x) \leq \alpha$ and from definition of $epi(f)$ we have $(x, \alpha) \in epi(f_i)$ and finally $(x, \alpha) \in \cap_{i \in I} epi(f_i)$.

(b) Let $(x, \alpha) \in X \times \mathbb{R}$ and $(x, \alpha) \in epi(min_{i \in I} f_i)$ we the same logic like (a) we conclude that $(x, \alpha) \in \cup_{i \in I} epi(f_i)$. ⊡

Example. The indicator function δ_C of a set $C \subset X$ is lower semicontinuous ⇔ C is closed.

Proof. Let a $\gamma \in \mathbb{R}$. The $\Gamma_\gamma(\delta_C)$ is the $\emptyset$ if $\gamma < 0$, and the set C otherwise. From the Theorem 2.4.8. we have the result. ⊡

Lemma 2.4.9. Let X be a Hausdorff space and let $(f_i)_{i \in I}$ be a family of lower-semicontinuous functions from X to the extended real line $\mathbb{R} \cup \{-\infty\} \cup \{+\infty\}$. Then $sup_{i \in I} f_i$ is lower-semicontinuous.

Proof. Since $epi(sup_{i\in I} f_i) = \cap_{i\in I}\, epi(f_i)$ and $epi(f_i)$ is a closed set and the intersection of closed sets are closed we have the result. ⊡

Theorem 2.4.10. Let (X, τ) be a Hausdorff space and let $f: X \to \mathbb{R} \cup \{+\infty\}$ be proper, lower-semicontinuous, and inf-compact. Then $argmin(f)$ is nonempty and compact. Moreover, $inf(f) \succ \infty$. ⊡

Minimizing Sequences

Definition 2.4.11. A function $f: X \to \mathbb{R} \cup \{+\infty\}$ is *sequentially lower-semicontinuous* at $x \in dom(f)$ if $f(x) \leq \lim_{n\to\infty} inf f(x_n)$ for every sequence x_n converging to x. ⊡

Definition 2.4.12. We say that x_n is a *minimizing sequence* for $f: X \to \mathbb{R} \cup \{+\infty\}$ if $\lim_{n\to\infty} f(x_n) = inf(f)$. ⊡

Proposition 2.4.13. Let x_n be a minimizing sequence for a function $f: X \to \mathbb{R} \cup \{+\infty\}$, which is sequentially l.s.c. and proper. If $x_n \to x$, then $x \in argmin(f)$. ⊡

Theorem 2.4.14. Let X be reflexive. If $f: X \to \mathbb{R} \cup \{+\infty\}$ is proper, convex, coercive ($\Gamma_\gamma(f)$is bounded $\forall \gamma \in$ R) and lower-semicontinuous, then $argmin(f)$ is nonempty and weakly compact. If moreover, f strictly convex, then $argmin(f)$ is a singleton. Therefore, theorem 2.4.14. assure us that for proper, convex, coercive, l.s.c. we have minimizers.

3 CONVEXITY

3.Introduction

In this section we discuss basic notions of convexity. First, we define convex sets. Convex sets help us to identify convex functions. We discuss about affine sets and we study separating theorems.

3.1. Convex Sets

Let x_1,$x_2 \in \mathbb{R}^n$, where $x_1 \neq x_2$. We define *line segment* between x_1 and x_2 points of the form

$$z = \lambda x_1 + (1-\lambda)x_2, \lambda \in \mathbb{R}.$$

We define *closed line segment* between x_1 and x_2 points of the form

$$z = \lambda x_1 + (1-\lambda)x_2, 0 \leq \lambda \leq 1.$$

We note that $z = x_2 + \lambda(x_1 - x_2)$, this means that z is the sum of the point x_2 and the direction $x_1 - x_2$ scaled by the parameter λ [4].

Let $C \subset \mathbb{R}^n$, we say that C is *convex* if $(1-\lambda)x + \lambda y \in C$ and $0 < \lambda < 1$. It means that a set $C \subset \mathbb{R}^n$ is convex if the line segment between any points in C lies in C. In particular, $\mathbb{H}$ and $\emptyset$ are convex [4]. ⊡

Theorem 3.1.1. The intersection of a collection of convex sets is convex. ⊡

A *convex combination* of $x_1, \dots, x_n \in C$ is a point of the form $\lambda_1 x_1 + \cdots + \lambda_m x_m$ where the coefficients $\lambda_i, i = 1, \dots, m$ is non-negative and $\sum_{i=1}^{m} \lambda_i = 1$.

Theorem 3.1.2. Let $C \subset \mathbb{R}^n$. The set C is convex if and only if it contains all convex combinations of its elements.

Proof. By definition C is convex $\Leftrightarrow \lambda_1 x_1 + \lambda_2 x_2 \in C$, $x_1, x_2 \in C, \lambda_1 \geq 0, \lambda_2 \geq 0$ and $\lambda_1 + \lambda_2 = 1$.

So for $m = 2$ the convexity it holds.

For $m > 2$ we suppose that C is closed under, taking all convex combination of fewer than m vectors:

Let $z = \lambda_1 x_1 + \cdots + \lambda_m x_m$ for $x_1, \ldots . x_m \in C$ and for some λ_i we have that $\lambda_i \neq 1$, otherwise $\sum_{i=1}^{m} \lambda_i = m \neq 1$. We choose arbitrary $\lambda_1 \neq 1$ and let $y = k_2 x_2 + \cdots + k_m x_m, k_m = \frac{\lambda_i}{1-\lambda_1}$. Then, $\sum_{i=2}^{m} k_i = \sum_{i=2}^{m} \lambda_i / \sum_{i=2}^{m} \lambda_i = 1$. After all y is a convex combination of $m-1$ elements of C , and from our hypothesis $y \in C$ and from the fact that $z = (1-\lambda_1)y + \lambda_1 x_1$ we have the result $x \in C$.[4] ⊡

The set of all convex combinations of points in C is called *convex hull* of C and is denoted by $\boldsymbol{conv}\ C$. In particular:

$$\boldsymbol{conv}\ C = \{\lambda_1 x_1 + \cdots + \lambda_m x_m | x_i \in C, \lambda_i \geq 0, \lambda_1 + \cdots + \lambda_m = 1\}. \quad \boxdot$$

It is obvious that convex hull is always a convex set. It is the smallest set that contains C. It is very interesting to obtain that the convex combination idea is useful in probability distributions. In general, let $C \subset \mathbb{R}^n$ a convex set and X is a random variable, where $X \in C$ with $pr = 1$, then $EX \in C$ [4]. ⊡

3.2. Affine Sets

Let $C \subset \mathbb{R}^n$, if for any $x_1, x_2 \in C$ and $\lambda \in \mathbb{R}$, we have $\lambda x_1 + (1-\lambda)x_2 \in C$, the set C is called *affine set*.

A *affine combination* of $x_1, \ldots, x_n \in C$ is a point of the form $\lambda_1 x_1 + \cdots + \lambda_m x_m$ where $\sum_{i=1}^{m} \lambda_i = 1$.

The set of all affine combinations in C is called *affine hull*, and is denoted by $\boldsymbol{aff}\ C$:

$$\boldsymbol{aff} C = \{\lambda_1 x_1 + \cdots + \lambda_m x_m | x_1, \ldots x_m \in C, \lambda_1 + \cdots + \lambda_m = 1\}.$$

The affine hull is the smallest set that contains C.The dimension of C,$dim(C)$ is the dimension of the $\boldsymbol{aff}\ C$. ⊡

Let $C \subset \mathbb{R}^n$. The *relative interior* of C is denoted as $ri(C)$ and is defined as:

$$ri(C) = \{x \in C | B(x,r) \cap \boldsymbol{aff} C \subseteq C \text{ for some } r > 0\}. \boxdot$$

Example 3.2.1. [4] We consider a square (x,y) −plane in $\mathbb{R}^3$: $C = z = (x,y,k) \in \mathbb{R}^3 | -1 \leq x \leq 1, -1 \leq y \leq 1$. The affine hull of C is $\boldsymbol{aff}\ C = \{z \in \mathbb{R}^3 | k = 0\}$. The $int(C) = \emptyset$ but the $ri(C) = \{z \in \mathbb{R}^3 \mid -1 < x < 1, -1 < y < 1, k = 0\}$. ⊡

We say that a function $f: \mathbb{R}^n \to \mathbb{R}^m$ is *affine* if it is a sum of a linear function and a constant.

3.3. Separating Theorems

Let $b \in \mathbb{R}^n, b \neq 0$ and any $\beta \in \mathbb{R}$. The sets:

- $\{x|\langle x, b\rangle \leq \beta\}$
- $\{x|\langle x, b\rangle \geq \beta\}$

Are called *closed half-spaces*. And the sets:

- $\{x|\langle x, b\rangle < \beta\}$
- $\{x|\langle x, b\rangle > \beta\}$

  Are called *open half-spaces*.

A hyperplane is a set of the form $\{\ x\ |\ \langle x - x_0, b\rangle\ = 0\ \}$. A hyperplane divides $\mathbb{R}^n$ into two half-spaces. Geometrically [4] the hyperplane is a set of points with a constant inner product to a vector b.

Theorem 3.3.1 (Hahn-Banach Separation Theorem). Let A and B be nonempty, disjoint convex of a normal space $(X, \|\ .\ \|)$.

(a) If A is open, there exist $K \in X^* \backslash \{0\}$ such that $\langle K, x\rangle < \langle K, y\rangle$ for each $x \in A$ and $y \in B$.

(b) If A is compact and B is closed, there exists $L \in X^* \backslash \{0\}$ and $\varepsilon > 0$ such that $\langle K, x\rangle + \varepsilon \leq \langle K, y\rangle \forall x \in A, y \in B$.

3.4 Convexity and Nonexpansiveness.

Let $T: \mathbb{H} \to \mathbb{H}$ an operator. We define the set of *fixed points* of T the set

$$FixT \coloneqq \{x \in \mathbb{H} :\ x = T(x)\}.$$

Non expansive operators are very useful, [2] because many optimization problems based to find fixed points of nonexpansive operators. Nonexpansive operators are Lipschitz continuous operators with $L = 1$.

Let $C \subset \mathbb{H}$, $C \neq \emptyset$ *and let* $T: C \to \mathbb{H}$. We say that T is:

(a) *Firmly nonexpansive,* if

$$\|Tx - Ty\|^2 + \|(I - T)x - (I - T)y\|^2 \leq \|x - y\|^2, \forall x, y \in C.$$

(b) *Nonexpansive,* if

$$\|Tx - Ty\| \leq \|x - y\|, \forall x, y \in C.$$

(c) *Contractive,* if

$$\|Tx - Ty\| \leq L\|x - y\|, L < 1, \forall x, y \in C.$$

It is obvious that statement (a) implies (b).

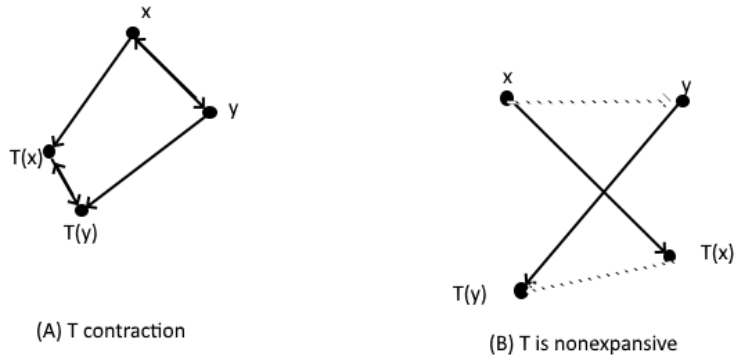


Figure 2 Geometrical Interpretation of contraction and nonexpansiveness

The interpretation of contraction is that mapping x, y to $T(x), T(y)$ reduces the distance between them and nonexpansive operator does not increase the distance between them [5].

Basic Properties.

- Let T_1, T_2 nonexpansive, then $T_1 \circ T_2$ is nonexpansive.
- Let T_1 a contraction and T_2 is nonexpansive, then $T_1 \circ T_2$ is contraction

Proposition 3.4.1. Let C nonempty set of $\mathbb{H}$. Let $T: C \to \mathbb{H}$. The T is firmly nonexpansive if, and only of, $I - T$ is firmly nonexpansive.

Let $C \subset \mathbb{H}$ a nonempty set. Let $T: C \to \mathbb{H}$ a nonexpansive operator and let a $\in (0,1)$. We say that T is *averaged* with constant a, or $a-$*averaged*, if there exists a nonexpansive operator $R: C \to \mathbb{H}$ such that $T = (1 - a)I + aR$. Note that if T is averaged, then is nonexpansive. By proposition 3.4.1. T is firmly nonexpansive if and only of is $\frac{1}{2}$-averaged.

Let $C \subset \mathbb{H}$ and let $(x_n) \in \mathbb{H}$. Then (x_n) is *Fejer monotone* with respect to C if

$$\forall x \in C \ \|x_{n+1} - x\| \leq \|x_n - x\|.$$

Proposition 3.4.2. Let $x_n \in \mathbb{H}$, and $C \subset \mathbb{H}$, $C \neq \emptyset$. If x_n Fejer monotone with respect to C. The we have the following:

(a) x_n is bounded.

(b) For every $x \in C$, $(\|x_n - x\|)_{n \in \mathbb{N}}$ converges.

Theorem 3.4.3. Let $x_n \in \mathbb{H}$, and $C \subset \mathbb{H}$, $C \neq \emptyset$. If x_n Fejer monotone with respect to C and that every weak sequential cluster point of $(x_n) \in C$. Then (x_n) converges weakly to a point $\hat{x} \in C$.

Krasnosel'skii-Mann Theorem

Theorem 3.4.4. Let $x_n \in \mathbb{H}$, and $C \subset \mathbb{H}$, $C \neq \emptyset$ and convex. Let $T: C \to C$ be a nonexpansive operator such that $FixT \neq \emptyset$, let $\lambda_n \in [0,1]$ such that $\sum_{n\in\mathbb{N}} \lambda_n(1-\lambda_n) = +\infty$, and let $x_0 \in C$. Set

$$\forall n \in \mathbb{N} \;\; x_{n+1} = x_n + \lambda_n(Tx_n - x_n)$$

Then the following statements are hold:

(a) x_n is Fejer monotone with respect to $FixT$.

(b) $(Tx_n - x_n)_{n\in\mathbb{N}}$ converges strongly to 0.

(c) x_n converges weakly to a point in $FixT$.

Proof. [2] (a)It holds the following corollary.

Corollary 3.4.5. Let $x \in \mathbb{H}$, $y \in \mathbb{H}$, and let $a \in \mathbb{R}$. Then

$$\|\alpha x + (1-\alpha)y\|^2 + \alpha(1-\alpha)\|x-y\|^2 = \alpha\|x\|^2 + (1-\alpha)\|y\|^2.$$

By corollary and definition of nonexpansiveness of T, we have for every $y \in FixT$

$$\|x_{n+1} - y\|^2 = \|(1-\lambda_n)(x_n - y) + \lambda_n(Tx_n - y)\|^2$$

$$= (1-\lambda_n)\|x_n - y\|^2 + \lambda_n\|Tx_n - Ty\|^2 - \lambda_n(1-\lambda_n)\|Tx_n - x_n\|^2$$

$$\leq \|x_n - y\|^2 - \lambda_n(1-\lambda_n)\|Tx_n - x_n\|^2.$$

This implies that (x_n) is Fejer monotone with respect to $FixT$.

(.b) From the last inequality we have $\sum_{n\in\mathbb{N}} \lambda_n(1-\lambda_n)\|Tx_n - x_n\|^2 \leq \|x_0 - y\|^2$. Since $\sum_{n\in\mathbb{N}} \lambda_n(1-\lambda_n) = +\infty$ we have $\underline{lim}\, \|Tx_n - x_n\| = 0$.

$$\|Tx_{n+1} - x_{n+1}\| = \|Tx_{n+1} - Tx_n + (1-\lambda_n)(Tx_n - x_n)\|$$

$$\leq \|x_{n+1} - x_n\| + (1-\lambda_n)\|Tx_n - x_n\|$$

$$= \|Tx_n - x_n\|.$$

This implies that $(Tx_n - x_n)_{n\in\mathbb{N}}$ converges strongly to 0.

(.c) Let x be a weak sequential cluster point of (x_n). Then from

Corollary 3.4.6. Let $D \subset \mathbb{H}$ closed, convex set. Let $T: D \to \mathbb{H}$ be nonexpansive, let $x_n \in D$, and let $x \in \mathbb{H}$. Suppose x a weak sequential cluster point of x_n and that $x_n - Tx_n \to 0 \Rightarrow x \in Fix(T)$.

Now apply theorem 3.4.3. and we have the result.

Proposition 3.4.7. Let $a \in (0,1)$, let $T: \mathbb{H} \to \mathbb{H}$ be an $a-$averaged operator such that $Fix(T) \neq \emptyset$, let $(\lambda_n)_{(n\in\mathrm{N})}$ be a sequence in $\left[0, \frac{1}{a}\right]$ such that $\sum_{n\in\mathbb{N}} \lambda_n(1-a\lambda_n) = +\infty$, and let $x_0 \in \mathbb{H}$. Set

$$x_{(n+1)} = x_n + \lambda_n(Tx_n - x_n), \forall n \in \mathrm{N}$$

Then the following hold:

(a) x_n is Fejer-Monotone w.r.t $Fix(T)$.

(b) $(Tx_n - x_n)_{(n\in\mathrm{N})}$ converges strongly to 0.

(c) (x_n) converges weakly to a point in $Fix(T)$.

Note. The previous theorems and propositions assure us the convergence of the algorithms that we will study in the next chapters.

4 CONVEX ANALYSIS AND SUBDIFFERENTIAL CALCULUS

4.Introfuction

In this chapter we define convex functions and their properties. We study the relation between convexity and continuity and convexity and differentiability. Then we characterize the convexity. In the second part we generalize the notion of derivative for nondifferentiable functions and will characterize their minimizers. We will discuss about proximal map and Moreau – Yosida Regularization, Legendre – Fenchel conjugate and finally about Fenchel - Rockafellar duality.

4.1. Convex Function

Let $f: X \to [-\infty, +\infty]$ be a function. The function f is *convex* if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \quad (3.1)$$

for each $x, y \in dom(f)$ and $\lambda \in (0,1)$.

This definition geometrically can be interpreted as the line segment between $(\alpha, f(\alpha))$, $(\beta, f(\beta))$, which is the chord from α to β, lies above the graph of f. Otherwise we can say that $f: X \to [-\infty, +\infty]$ is convex if and only if its $epi(f)$ convex [4].

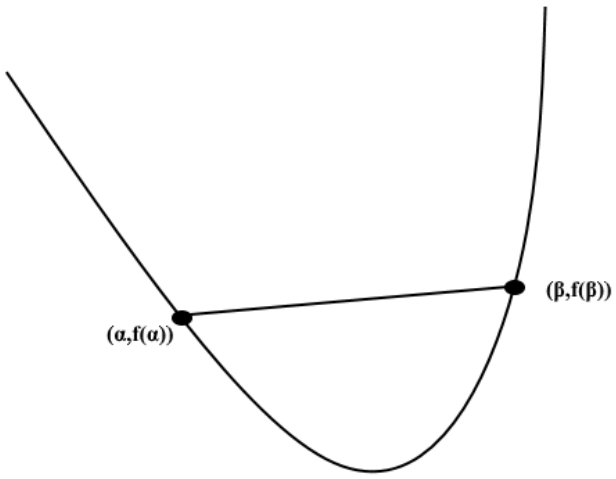


Figure 3 Convex function.

The function f is *strictly convex* if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$$

for each $x, y \in dom(f)$ and $\lambda \in (0,1)$.

The function f is *strongly convex* with parameter $\mu > 0$ if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \frac{\alpha}{2}\lambda(1-\lambda)\|x-y\|^2$$

for each $x, y \in dom(f)$ and $\lambda \in (0,1)$.The inequality (3.1) is called Jensen's inequality and it is extended to convex combinations for m points, where $m > 2$, so we have :

$$f(\lambda_1 x_1 + \cdots + \lambda_m x_m) \leq \lambda_1 f(x_1) + \cdots + \lambda_m f(x_m).$$

We can observe that if f is convex, then each γ −sublevel set is convex. We say that f is *concave* if $-f$ is convex. In the same way, *strictly concave*.

Example 4.1.1. We suppose the indicator function δ_C , where C is convex set, then δ_C is a convex function.

Example 4.1.2. Let f be a function on $\mathbb{R}^n$.

(a) If f is a norm, then it is a convex function.

Proof. $f(\theta x + (1-\theta)y) = \|\theta x + (1-\theta)y\|$

$$\leq \|\theta x\| + \|(1-\theta)y\| \text{ (triangle inequality)}$$
$$= \theta\|x\| + (1-\theta)\|y\|$$
$$= \theta f(x) + (1-\theta)f(y).$$

(b) If $f(x) = max\{x_1, \dots x_n\}$ then is convex.

(c) If f is the Tchebycheff norm , $f(x) = max|k_i|, i = 1, \dots n$, is convex function.

The *support function* $\delta^*(.|C)$ of a convex set $C \subset \mathbb{R}^n$ is:

$$\delta^*(.|C) = sup\{\langle x, y\rangle | y \in \lambda C\}.$$

Theorem 4.1.3. If f_1 and f_2 are proper convex function on $\mathbb{R}^n$, then $f_1 + f_2$ is convex.

Proof. Indeed, from the definition of convex function it is elementary.

Theorem 4.1.4. The pointwise supremum of an arbitrary collection of convex functions is convex.

Proof. As we know, the intersection of a collection of convex sets is convex. We have $f(x) = sup\{fi(x)|i \in I\}$, where f_i are convex functions for each i. Indeed, the $epi(f)$.

We define the *lower semicontinuous hull* of f :

$$(clf)(x) = \begin{cases} liminf_{y \to x} f(y), if \quad f(y) > \infty \quad for\ all\ x \in X \\ -\infty, otherwise \end{cases}$$

We say that function f is *closed* if $f = clf$. The closedness is equivalent with lower-semicontinuity.

4.2. Convexity and continuity

This subsection is following to [1].

Proposition 4.2.1. Let $f: X \to \mathbb{R} \cup \{+\infty\}$ proper function. Then, f is convex and lower semicontinuous if, and only if, there exists a family of functions $(f_i)_{i\in I}$ if continuous affine functions on X such that $f = sup(f_i)$.

Characterization of Continuity

Proposition 4.2.1. [1] Let $f: X \to \mathbb{R} \cup \{+\infty\}$ a convex function and a point $x_0 \in X$. We have the following equivalent statements:

(a) f is upper bounded in a $V(x_0)$

(b) f is Lipschitz-continuous in a $V(x_0)$

(c) f is continuous in $x_0 \in X$

(d) $(x_0, a) \in int\big(epi(f)\big)$ for each $\lambda > f(x_0)$.

Note. Let $(X, \|\,.\,\|)$ be a normed space. Let $f: X \to \mathbb{R} \cup \{+\infty\}$ a convex function. We know that f is continuous in $int\big(dom(f)\big)$ in the next three cases:

(i) X is finite dimensional.

(ii) X is Banach space and f is l.s.c.

(iii) f is continuous at a point x.

4.3. Convexity and Differentiability.

Let a function $f: X \to [-\infty, +\infty]$. We define the *directional derivative* of function f at a point x in $domain$ of f, $dom(f)$ in the direction h the quantity:

$$f'(x; h) = \lim_{t\to 0} \frac{f(x+th)-f(x)}{t}.$$

We define the one-sided directional derivative of f at $x \in X$ to the direction h to be the limit $f'(x; h) = \lim_{t\downarrow 0} \frac{f(x+th)-f(x)}{t}$.

One of the most useful property of convex functions is that the one-sided directional derivative is always exists in $\mathbb{R} \cup \{+\infty\}$.

Theorem 4.3.1. Let f convex function and let a point x such that $f(x) < +\infty$. For each h, the difference quotient in the definition of $f'(x;h)$ is a non-decreasing function of $t > 0$,so that $f'(x,h)$ exists and $f'(x;h) = \lim_{t>0} \frac{f(x+th)-f(x)}{t}$.

Proof. [13] The difference quotient for $t > 0$ can be expressed as $t^{-1}g(th)$, where $g(h) = f(x+h) - f(x)$. The set $epi(g)$ and can be interpreted as the removal of point $\big(x, f(x)\big)$ to (0,0). Also, $t^{-1}g(th) = (gt^{-1})(h)$. From the fact that $epi(g)$ is convex, we have that also the set $t^{-1}epi(g)$ is convex, so the function gt^{-1} is convex. Since $epi(g)$ contains the origin, the latter set increases, as t^{-1} decreases. ⊡

Proposition 4.3.2. Let $f: X \to \mathbb{R} \cup \{+\infty\}$ proper and convex, and let $x \in dom(f)$. We define the function $\varphi_x: X \to [-\infty, +\infty]$, as $\varphi_x(h) = f'(x,h)$. The function $\varphi_X(h)$ is convex and if f is continuous in x, then φ_x is finite and continuous in X.⊡

If the above function φ_x is linear and continuous in X, in a point $x \in dom(f)$ we say that the function f is $Gateaux - differentiable$ (GD) at x. The *Gâteaux derivative* or *gradient* of f at x is $\nabla f(x) = f'(x;.)$ and $\nabla f(x) \in X^*$.

A function f is *Fréchet-differentiable* at x if there exists $L \in X^*$ such that

$$lim_{\|h\|\to 0} \frac{|f(x+h) - f(x) - \langle L, h\rangle|}{\|h\|} = 0.$$

The *Fréchet derivative* of f at x is $Df(x) = L$.

Proposition 4.3.3. (Descent Lemma). If $f: X \to \mathbb{R} \cup \{+\infty\}$ is Gateux-Differentiable and ∇f is Lipschitz – continuous with constant L, then

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2$$

for each $x, y \in X$. In particular, f is continuous.

Proof. The proof is according [1]. Let $h = y - x$ and define $g: [0,1] \to \mathbb{R}$ by $g(t) = f(x + th)$. Then $\dot{g}(t) = \langle \nabla f(x + th), h\rangle$ for each $t \in (0,1)$, and so

$$\int_0^1 \langle \nabla f(x + th), h\rangle dt = \int_0^1 \dot{g}(t)dt = g(1) - g(0) = f(y) - f(x).$$

Therefore,

$$f(y) - f(x) = \int_0^1 \langle \nabla f(x), h\rangle\, dt + \int_0^1 \langle \nabla f(x + th) - \nabla f(x), h\rangle\, dt$$

$$\le \langle \nabla f(x), h\rangle + \int_0^1 \|\nabla f(x + th) - \nabla f(x)\|\|h\| dt$$

$$\leq \langle \nabla f(x), h \rangle + L\|h\|^2 \int_0^1 t dt$$

$$= \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2 . \square$$

Let $f: X \to \mathbb{R}$ is GD in X. The *directional derivative of* $\nabla f: X \to X^*$ is the function

$$(\nabla f)'(x; h) = lim_{t \to 0^+} \frac{\nabla f(x+th) - \nabla f}{t}.$$

The function f is *twice Gâteaux-differentiable* if is Gâteaux differentiable and $(\nabla f)'(x; h)$ exists for all $h \in X$, and the function $h \mapsto (\nabla f)'(x; h)$ is linear and continuous. The second Gâteaux derivative (Hessian) of f at $x \in X$ is $\nabla^2 f(x) = (\nabla f)'(x, .) \in \mathcal{J}(X; X^*) . \square$

Characterization of Convexity [1]

Theorem 4.3.4. (Fermat's Rule). Let a normed space $(X, \| . \|)$ and $C \subset X$ convex set. Let $f: X \to \mathbb{R} \cup +\{\infty\}$. If $f(x) \leq f(y)$ for all $y \in C$ and if f is Gateaux differentiable at x, then

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

for all $y \in C$. If moreover $x \in int(C)$, then $\nabla f(x) = 0$.

Proof. Let $y \in C$, from convexity of C we have

$$z = \lambda y + (1 - \lambda)x \in C \text{ for } \lambda \in (0,1).$$

The inequality $f(x) \leq f(z) \Leftrightarrow f\big(x + \lambda(y - x)\big) - f(x) \geq 0$. If we divide by λ the last inequality and let, $\lambda \to 0$ we have $f'(x; y - x) \geq 0$ for all $y \in C$. $\square$

To understand the Fermat's Rule, [1] let f a differentiable function on $\mathbb{R}^2$. The Theorem means that f decrease by leaving the set C.

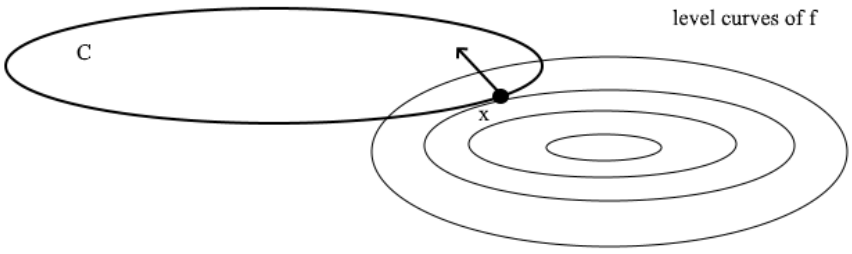


Figure 4 Fermat's rule. The vector is the gradient of f

We conclude that Fermat's rule gives us a necessary condition for a point $\hat{x}$ be a minimizer of f. We have the following

$$\hat{x} \text{ is minimizer of } f \Leftrightarrow \nabla f(\hat{x}) = 0.$$

Proposition 4.3.5. Let $f: C \to \mathbb{R}$ be Gateaux-differentiable, where $C \subset X$ is convex and open set. The convexity is characterized by the equivalent statements:

(a) f is convex.

(b) $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$, for every $x, y \in C$.

(c) $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0$, for every $x, y \in C$.

If f is twice GD on C,

(d) $\langle \nabla^2 f(x)h, h \rangle \geq 0$, for every $x \in C$ and $h \in X$. (*positive semidefinite*)

Proof. The proof is according to [1]

By convexity of f we have for all $y \in X$ and $\lambda \in (0,1)$,

$$f(\lambda y + (1-\lambda)x) \leq \lambda f(y) + (1-\lambda) f(x)$$

$$\Leftrightarrow \frac{f(\lambda y + (1-\lambda)x)}{\lambda} \leq f(y) - f(x).$$

For $\lambda \to 0$ we obtain b). From b) we have obvious the inequality c).

c) ⇒ a) Let $g: [0,1] \to \mathbb{R}$, where $g(\lambda) = f(\lambda x + (1-\lambda)y) - \lambda f(x) - (1-\lambda) f(y)$.

We obtain $g(0) = g(1) = 0$ and

$$g'(\lambda) = \langle \nabla f(\lambda x + (1-\lambda)y), x - y \rangle - f(x) + f(y)$$

For $\lambda \in (0,1)$. Take $0 < \lambda_1 < \lambda_2 < 1$ and write $x_i = \lambda_i x + (1-\lambda_i)y$ for $i = 1,2$.

$$g'(\lambda_1) - g'(\lambda_2) = \frac{1}{\lambda_1 - \lambda_2} \langle \nabla f(x_1) - \nabla f(x_2), x_1 - x_2 \rangle \leq 0$$

This implies that g' is nondecreasing. Since $g(0) = g(1) = 0$, there exists $\xi \in (0,1)$ such that $g(\xi) = 0$. Since

- g' nonincreasing
- $g' \leq 0$ on $[0, \xi]$
- $g' \geq 0$

We have that $g(\lambda) \geq 0$ and f convex.

d)⇒c)⇒a) We assume that f is twice GD. Let $t > 0$ and $h \in X$, we have $\langle \nabla f(x + th) - \nabla f(x), th \rangle \geq 0$. Now,

- We divide by t^2.
- We take the limit as $t \to 0$.

We have $\langle \nabla^2 f(x)h, h \rangle \geq 0$. Finally,

$$g''(\lambda) = \langle \nabla^2 f(\lambda x + (1-\lambda)y)(x - y), x - y \rangle \geq 0.$$

It follows that g' is nonincreasing and we conclude like before. ⊡

The strict convexity characterized as in proposition 4.3.5. but the inequalities are hold strict. Let $f: C \to \mathbb{R}$ be GD, where $C \subset X$ is open and convex then the following statements are equivalent:

(a) f is strictly convex

(b) $f(y) > f(x) + \langle \nabla f, y - x \rangle$, for any $x \neq y \in C$.

(c) $\langle \nabla f(x) - \nabla f(y), x - y \rangle > 0$, for any $x \neq y \in C$.

If additionally f is twice GD on C, then the following is equivalent with the previous:

(d) $\langle \nabla^2 f(x)h, h \rangle > 0$, for every $x \in C$ and $h \in X$.

[1] (Characterization of strong convexity). Let $C \subset X$ be open and convex, and let $f: C \to \mathbb{R}$ be GD. The following are equivalent.

(a) f is a −strongly convex

(b) $f(y) > f(x) + \langle \nabla f, y - x \rangle + \frac{\alpha}{2}\|x - y\|^2$, for any $x, y \in C$.

(c) $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq a\|x - y\|^2$, for any $x, y \in C$.

If moreover, f is twice GD on C, then the following is equivalent with the previous:

(d) $\langle \nabla^2 f(x)h, h \rangle \geq \frac{\alpha}{2}\|h\|^2$, for every $x \in C$ and $h \in X$.

Geometrical interpretation of convex differentiable function is that the hyperplane

$$H = \{(y, z) \in X \times \mathbb{R}: f(x) + \langle \nabla f(x), y - x \rangle = z\}$$

lies below the epigraph of f, $epi(f)$ and touches it and point $\big(x, f(x)\big)$. In other words, $\nabla f(x)$ is a non-vertical supporting hyperplane of $epi(f)$ at $\big(x, f(x)\big)$.

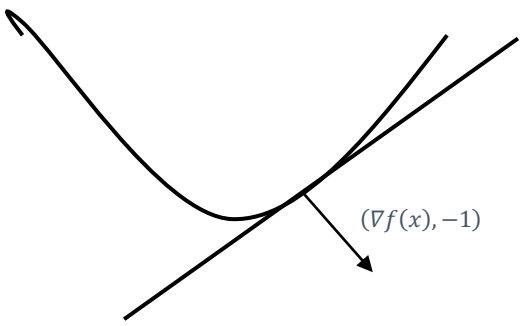


Figure 5 Geometrical Interpretation of convex differentiable function.

4.4. Subgradients

The idea of subgradients is to generalize the notion of gradient ∇f to non-differentiable function. We can generalize the convex inequality

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

for a function f, where f is not necessarily at x.

Let a function $f: X \to [-\infty, +\infty]$, convex and lower-semicontinuous. A vector $x^* \in X^*$ is a *subgradient* of function f at point x if

$$f(z) \geq f(x) + \langle x^*, z - x \rangle, \forall z.$$

The set of all subgradients at x is called the s*ubdifferential* of f, is denoted by ∂f and is defined:

$$\partial f(x) = \{x^* \in X^* | f(y) \geq f(x) + \langle x^*, y - x \rangle, \text{for all y} \in \text{X}\}.$$

We say that the function f is *subdifferentiable* at a point x if $\partial f(x) \neq \emptyset$.The *domain* of ∂f is the set: $dom(\partial f) = \{x \in X | \partial f(x) \neq 0\}$. Is is obvious that, $dom(\partial f) \subset dom(f)$.

Geometrical Interpretation of Subgradients

Let $x \in X$.We assume that f is finite at x. We assume the function

$$g(z) = f(x) + \langle x^*, z - x \rangle.$$

Recall that we as function is affine if it is a sum of a linear function and a constant. The function $g(x)$ is affine and is a non-vertical supporting hyperplane to the convex set $epi(f)$ at the point $\big(x, f(x)\big)$.

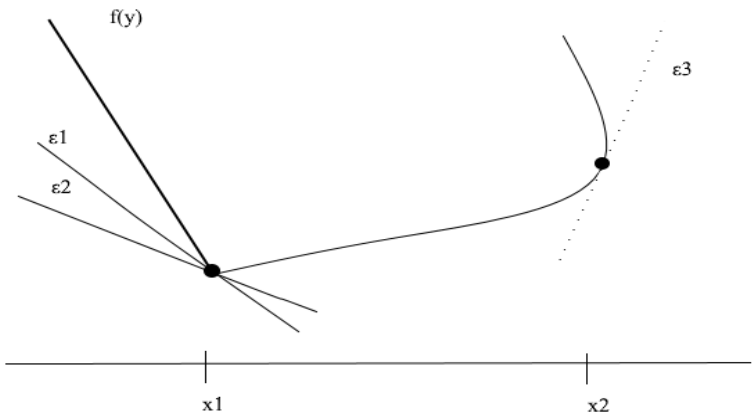


Figure 6 Geometrical interpretation of subgradients

The subgradient gives affine global underestimator of f.

Properties of Subdifferential ∂f

It is obvious that the subdifferential of f at x is a closed convex set. Since it is the intersection of closed convex half-spaces [13] $H = \{x^* \mid f(z) \geq f(x) + \langle x^*, x - z\rangle\}$, and the intersection of closed, convex set is a closed convex set.

Proposition 4.4.1. The set $\partial f(x)$ is closed and convex, $\forall x \in X$.

Proof. [1] For convexity.

Let $x_1^*, x_2^* \in \partial f(x)$ and $t \in (0,1)$. For each $z \in X$ and from the definition of subgradient, we have:

- $f(z) \geq f(x) + \langle x_1^*, z - x\rangle \quad (1)$
- $f(z) \geq f(x) + \langle x_2^*, z - x\rangle \quad (2)$

If we add t times the (1) inequality and $1 - t$ times the inequality (2), we have

$$tf(z) \geq tf(x) + t\langle x_1^*, z - x\rangle \Leftrightarrow tf(z) \geq tf(x) + \langle tx_1^*, z - x\rangle \ (3)$$

$$(1-t)f(z) \geq (1-t)f(x) + \langle (1-t)x_2^*, z - x\rangle \ (4)$$

If we add (3) and (4) we have

$$f(z) \geq f(x) + \langle tx_1^* + (1-t)x_2^*, z - x\rangle \Leftrightarrow tx_1^* + (1-t)x_2^* \in \partial f(x)$$

For the closed.

We take a sequence $x_n^* \in \partial f(x)$, where $x_n^* \to x^*$. Since, $x_n^* \in \partial f(x)$ we have

$$f(z) \geq f(x) + \langle x_n^*, z - x\rangle, \forall z \in X \text{ and } n \in \mathbb{N}.$$

Let $n \to \infty$ we have

$$f(z) \geq f(x) + \langle x^*, z - x\rangle \Leftrightarrow x^* \in \partial f(x).$$

Examples 4.4.2.

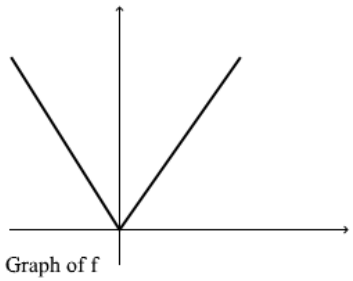
1) Let $f: \mathbb{R} \to \mathbb{R}$, $f(x) = |x|$. The function of absolute value is differentiable at every $x \neq 0$. Let's calculate the subgradient at $x = 0$.

$$\partial f(0) = \{x^* \mid f(y) \geq f(0) + \langle x^*, y - 0\rangle\} =$$

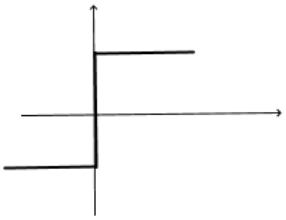
$$= \{\, x^* \mid |y| \geq \langle x^*, y\rangle \,\}$$

$$= \{\, x^* \mid |y| \geq x^* y \,\}$$

$$= [-1,1].$$



Graph of f



Graph of sub f

2) Let $f: X \to \mathbb{R}$, $f(x) = \|x\|_2$ the Euclidean norm. It is subdifferential at every $x \in X$ and differentiable at every $x \neq 0$. The subgradient is:

$$\partial f(0) = \{\, x^* \mid \| x^* \| \leq 1 \,\} = B_X(0,1).$$

3) Let $f: X \to \mathbb{R}$, $f(x) = \|x\| = max\{s^T x, s_i \in \{-1, +1\}\}$. We have

$$\partial f(0,0) = [-1,1] \times [-1,1],$$
$$\partial f(1,0) = 1 \times [-1,1],$$
$$\partial f(1,1) = \{(1,1)\}.$$

4) Let $C \subset X, C \neq \emptyset$ closed and convex set. Let $\delta_C: X \to \mathbb{R} \cup +\infty$, the indicator function, we have:

$$z \in \partial \delta_C(x) \Leftrightarrow \delta_C(y) \geq \delta_C(x) + \langle z, y - x \rangle \; \forall y.$$

It follows that $\partial \delta_C(x)$ is the *normal cone* to C at x.

In this part we will analyze some very useful propositions.

Proposition 4.4.3. Let $f: X \to \mathbb{R} \cup \{+\infty\}$ be convex. If f is GD at a point x, then $x \in dom(\partial f)$ and $\partial f(x) = \{\nabla f(x)\}$.

Proof. [1] From the convexity of f we have the inequality

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

and from the subgradient inequality we can imply $\nabla f(x) \in \partial f(x)$.

Let $x^* \in \partial f(x)$. We must prove that x^* is unique and necessarily $x^* = \nabla f(x)$.By definition,

$$f(y) \geq f(x) + \langle x^*, y - x \rangle \; \forall y \in X.$$

Take any $h \in X$ and $t > 0$, and write $y = x + th$, and from above inequality we have

$$\frac{f(x+th)-f(x)}{t} \geq \langle x^*, h \rangle.$$

If we take the limit as $t \to 0$, we have,

$$\langle \nabla f(x) - x^*, h \rangle \geq 0 \; \forall h \in X.$$

Therefore, $x^* = \nabla f(x)$.

Proposition 4.4.5. Let a convex function, $f: X \to \mathbb{R} \cup \{+\infty\}$ for $x^* \in \partial f(x)$ and $y^* \in \partial f(y)$, then $\langle x^* - y^*, x - y \rangle \geq 0$.

Proof. We have,

$$x^* \in \partial f(x) \Leftrightarrow \mathrm{f(y)} \geq \mathrm{f(x)} + \langle \mathrm{x^*, y - x} \rangle \; (1)$$

$$y^* \in \partial f(y) \Leftrightarrow \mathrm{f(x)} \geq \mathrm{f(y)} + \langle \mathrm{y^*, x - y} \rangle \; (2)$$

If we add (1) and (2) we have,

$\mathrm{f(y)} + \mathrm{f(x)} \geq \mathrm{f(x)} + \mathrm{f(y)} + \langle \mathrm{x^*, y - x} \rangle + \langle \mathrm{y^*, x - y} \rangle \Leftrightarrow \langle x^* - y^*. x - y \rangle \geq 0.$

With the previous proposition we generalize the non-decreasing monotonicity of a differentiable function. The subgradient ∂f is a monotone operator. Also, we can generalize the *Fermat's Rule*.

Theorem 4.4.6. Let $f: X \to \mathbb{R} \cup \{+\infty\}$ a proper and convex function. The element $\hat{x}$ is a global minimizer of $f \Leftrightarrow 0 \in \partial f(\hat{x})$.

Proof. Let $g = 0$ be a subgradient of f at $x^* \Rightarrow f(y) \geq f(x^*) + 0 \Rightarrow f(y) \geq f(x^*) \Rightarrow$ x^* is global minimizer of f. And the opposite direction, let $\hat{x}$ be a global minimizer of f, then $f(x) \geq f(\hat{x}) \Leftrightarrow f(x) \geq f(\hat{x}) + \langle 0, x - \hat{x} \rangle \Leftrightarrow 0 \in \partial f(\hat{x})$.

The Fermat's rule is sufficient condition for $\hat{x}$ be a global minimizer of f.

Proposition 4.4.5. Let a convex function $f: X \to \mathbb{R} \cup \{+\infty\}$ and continuous at x, then $\partial f(x)$ is bounded and $\partial f(x) \neq \emptyset$.

The converse of *proposition 4.4.5* it is not true. For example [1], let $f: \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$, $f(x) = +\infty$ if $x \neq 0$, and $f(0) = 0$. Then the $\partial f(x) = \begin{cases} \emptyset, if\ x \neq 0 \\ +\infty, otherwise \end{cases}$. It follows that function f is subdifferentiable but not continuous at 0.

4.5. Subdifferential Calculus

Sum of convex functions

In this section we refer a basic and very useful theorem. The Moreau-Rockafellar theorem. This theorem is about the relation between the subgradient of the sum of two convex function and the sum of subgradients of two functions. Theory on this subsection helps us to define duality (next chapter) and to find minimizer for convex functions, more things will discuss in the next subsection.

Theorem 4.5.1. Let $f, g: X \to \mathbb{R} \cup \{+\infty\}$ be proper, convex, lower semicontinuous. $\forall x \in X$ we have,

$$\partial f(x) + \partial g(x) \subset \partial(f + g)(x).$$

If f is continuous at some $x_0 \in dom(g)$, then $\partial f(x) + \partial g(x) = \partial(f + g)(x) \forall x \in X$

Proof. [1] We take $x^* \in \partial f(x)$ and $y^* \in \partial g(x)$, then

$$f(y) \geq f(x) + \langle x^*, y - x \rangle \text{ and } g(y) \geq g(x) + \langle y^*, y - x \rangle \ \forall y \in X.$$

If we add the two inequalities, we have

$f(y) + g(y) \geq f(x) + g(x) + \langle x^* + y^*, y - x \rangle \ \forall y \in X,$

The last inequality implies that $x^* + y^* \in \partial(f + g)(x)$.

We take $u^* \in \partial(f + g)(x)$. We have

$$g(y) + g(y) \geq f(x) + g(x) + \langle u^*, y - x \rangle \text{ for every } y \in X.$$

We need to find $x^* \in \partial f(x)$ and $y^* \in \partial g(x)$ such that $x^* + y^* = u^*$. We define the convex nonempty sets:

$B = \{(y, \lambda) \in X \times \mathbb{R}: g(y) - g(x) \leq -\lambda\}$

$C = \{(y, \lambda) \in X \times \mathbb{R}: f(y) - f(x) - \langle u^*, y - x \rangle \leq \lambda\}$ and,

$h: X \to \mathbb{R} \cup +\infty$ as $h(y) = f(y) - f(x) - \langle u^*, y - x \rangle$, h is continuous in x_0 and $C = epi(h)$, the open convex set $A = int(C)$ is nonempty from proposition (char of continuity) and the inequality

$$g(y) + g(y) \geq f(x) + g(x) + \langle u^*, y - x \rangle$$

We have $A \cap B = \emptyset$ and from Hahn Banach theorem we obtain a $(K, s) \in X^* \times \mathbb{R} \backslash \{(0,0)\}$ such that

$$\langle K, y \rangle + s\lambda \leq \langle K, z \rangle + s\mu, \forall (y, \lambda) \in A, (z, \mu) \in B.$$

We take $(y, \lambda) = (x, 1) \in A$ and $(z, \mu) \in B$, we conclude that $s \leq 0$.

If we take $s = 0$ and $z = x_0$ we have that $\langle K, x_0 - y \rangle \geq 0 \forall y \in V(x_0)$ and it follows K=0 and it is a contradiction to $(K, s) \neq (0,0)$. Therefore $s < 0$. For $y^* = -\frac{L}{s}$ we have

$$\langle y^*, y \rangle + \lambda \leq \langle y^*, z \rangle + \mu.$$

By the definition of C, we take $(z, \mu) = (x, 0) \in B$ and we have

$$\langle y^*, y - x \rangle + f(y) - f(x) - \langle y^*, y - x \rangle \leq 0.$$

From inequality $g(y) + g(y) \geq f(x) + g(x) + \langle u^*, y - x \rangle$ we have

$$f(z) \geq f(x) + \langle u^* - y^*, z - x \rangle \forall z \in X,$$

therefore, $x^* = u^* - y^* \in \partial f(x) \forall x \in X$. ⊡

Note. If f is continuous at some $x_0 \in dom(g)$, we have

$$\partial f(x) + \partial g(x) = \partial (f + g)(x), \forall x \in X \Rightarrow$$

$$dom\big(\partial (f + g)\big) = dom(\partial f) \cap dom(\partial g).$$

Chain Rule. Let $A \in X^*$ and let $f: Y \to \mathbb{R} \cup \{+\infty\}$ be proper, convex, and lower-semicontinuous. For each $x \in X$, we have

$$A^* \, \partial f(Ax) \subset \partial (f \circ A)(x)$$

If f is continuous at some $y_0 \in A(X)$,we have the equality,

$$A^* \, \partial f(Ax) = \partial (f \circ A)(x)$$

From Chain Rule and Moreau – Rockafellar theorem we can conclude,

$$A^* \partial f(Ax) + \partial g(X) \subset \partial(f \circ A + g)(x),$$

for $A \in X^*$ and two functions f, g proper, convex and lower-semicontinuous.

If there is $x_0 \in dom(g)$ such that f is continuous at Ax_0, then

$$A^* \partial f(Ax) + \partial g(X) \subset \partial(f \circ A + g)(x).$$

4.6. Proximal Map and Moreau - Yosida Regularization.

In this subsection we define functions on Hilbert Space $\mathbb{H}$. In convex optimization is very common to minimize convex function, which is not smooth, like the $\ell_1 -$ norm, the TV-debluring, or least squares. We need to find a way to handle these functions. The idea is to create a smooth version of the non-smooth function. We success smoothness by adding a quadratic term.

We define *Moreau-Yosida Regularization* of f with parameter (λ, x), for a given $\lambda > 0$ and $x \in \mathbb{H}$ the function $f_\lambda(x) = min_{x \in \mathbb{H}} f(x) + \frac{1}{2\lambda}\|x - z\|^2$. The function f_λ is a smooth function $\forall\, \lambda > 0$.

Proposition 4.6.1. For each $\lambda > 0$ and $x \in \mathbb{H}$, the function,

$$z \longmapsto f_{(\lambda x)}(z) \coloneqq f(z) + \frac{1}{2\lambda}\|x - z\|^2 ,$$

has a unique minimizer $\hat{x}$ and is characterized by the relation,

$$-\frac{\hat{x}-x}{\lambda} \in \partial f(\hat{x}).$$

Proof. [1] The function $f_{(\lambda,x)}$ is proper, convex and l.s.c. but also is strictly convex and coercive, because f is proper, convex and lower-semicontinuous. Therefore, from Theorem 2.4.15 we know that $f_{(\lambda,x)}$ has a unique minimizer $\hat{x}$. From the Fermat's Rule 4.4.6 the unique minimizer $\hat{x}$ satisfies the optimally condition and the Moreau – Rockafellar Theorem 4.5.1. we have

$$0 \in \partial f_{(\lambda,x)}(\hat{x}) = \partial f(\hat{x}) + \frac{\hat{x}-x}{\lambda} \Leftrightarrow -\frac{\hat{x}-x}{\lambda} \in \partial f(\hat{x}). \boxdot$$

After all, if f is convex, proper and l.s.c. ,then , for any x, there is a unique minimizer $\hat{x}$ to the strongly convex problem $argmin_{x \in \mathbb{H}} f(z) + \frac{1}{2\lambda}\|x - z\|^2$. We define

$$\hat{x} =: prox_{\lambda f}(x),$$

and is called *proximity* or *proximal operator* of f.

In general, we define *resolvent* of a monotone operator T, the quantity $(I + \lambda T)^{-1}$, where I the identity relation. As we prove in proposition 4.4.5. the ∂f is monotone operator and we can define the *proximal operator* by $prox_{\lambda f} = (I + \lambda\, \partial f)^{-1}$, as the resolvent of subgradient ∂f.

Proposition 4.6.2. For a proper, convex and lower-semicontinuous function, $f: \mathbb{H} \to \mathbb{R} \cup \{+\infty\}$ the proximal operator $prox_{\lambda f}: \mathbb{H} \to \mathbb{H}$ is nonexpansive operator.

Proof. [1] Let $\hat{x} = prox_{\lambda f}(x)$ and $\hat{y} = prox_{\lambda f}(y)$, so from the previous proposition 4.6.1 we have,

$$-\frac{\hat{x}-x}{\lambda} \in \partial f(\hat{x}) \text{ and } -\frac{\hat{y}-y}{\lambda} \in \partial f(\hat{y}).$$

Since ∂f is monotone, we have

$$\langle (\hat{x} - x) - (\hat{y} - y), \hat{x} - \hat{y} \rangle \leq 0.$$

This implies,

$$0 \leq \|\hat{x} - \hat{y}\|^2 \leq \langle x - y, \hat{x} - \hat{y} \rangle \leq \|x - y\| \|\hat{x} - \hat{y}\|$$

and therefore,

$$\|\hat{x} - \hat{y}\| \leq \|x - y\|. \ \square$$

Proposition 4.6.3. For proper closed convex function f and $\lambda > 0$, $prox_{\lambda f}$ is firmly nonexpansive.

Proof. Similar with the above proposition.

The notion of firm nonexpansive is very useful for the convergence of proximal algorithms, as we shall discuss in the next chapter.

We obtain that the proximal operator in Hilbert space is the unique point $\hat{x}$ [2] which satisfies

$$f_\lambda(x) = f(\hat{x}) + \frac{1}{2\lambda}\|x - \hat{x}\|^2$$

Now we will prove according to [2], that the fixed points of a proximal operator are the minimizers of f. This is useful, as in algorithms we will minimize convex functions finding fixed point of nonexpansive operators.

Proposition 4.6.4. Let f a proper, lower-semicontinuous convex function on $\mathbb{H}$ to extended real line and let $x, p \in \mathbb{H}$. Then

$$p = prox_f(x) \Leftrightarrow \forall y \in \mathbb{H}\ \langle y - p, x - p\rangle + f(p) \leq f(y)$$

Proof. [2] Let $y \in \mathbb{H}$. We suppose $p = prox_f$ and for each $a \in (0,1)$, $z = ay + (1-a)p$. For every $a \in (0,1)$ from definition of proximal operator and the convexity of f we have

$$f(p) \leq f(z) + \frac{1}{2}\|x - z\|^2 - \frac{1}{2}\|x - p\|^2$$

$$\leq af(y) + (1-a)f(p) - a\langle x - p, y - p\rangle + \frac{a^2}{2}\|y - p\|^2$$

$$\Leftrightarrow \langle y - p, x - p\rangle + f(p) \leq f(y) + \tfrac{a^2}{2}\|y - p\|^2.$$

Letting $a \to 0$, we have the inequality.

We suppose now that $\langle y - p, x - p\rangle + f(p) \leq f(y)$ then

$$f(p) + \frac{1}{2}\|x - y\|^2 \leq f(y) + \frac{1}{2}\|x - p\|^2 + \langle x - p, p - y\rangle + \frac{1}{2}\|p - y\|^2$$
$$= f(y) + \frac{1}{2}\|x - y\|^2$$

and this implies $p = prox_f$. ⊡

Proposition 4.6.5. Let f proper, lower-semicontinuous convex function on $\mathbb{H}$ to extended real line. Then

$$Fix\big(prox_f\big) = Argmin(f).$$

Proof. [2] Let $x \in \mathbb{H}$. Then from proposition 4.6.4. for

$$x = prox_f(x)$$

$$\Leftrightarrow \forall y \in \mathbb{H}\ \langle y - x, x - x\rangle + f(x) \leq f(y)$$

$$\Leftrightarrow \forall y \in \mathbb{H}\ f(x) \leq f(y)$$

$$\Leftrightarrow x \in argmin(f).\ \boxdot$$

4.7. The Legendre – Fenchel conjugate

Let a function $f: X \to \mathbb{R} \cup \{+\infty\}$ proper, we define the *Legendre – Fenchel conjugate* (or convex conjugate) be the function $f^*: X^* \to \cup +\infty$,

$$f^*(x^*) = sup_{x \in X}\{\langle x^*, x\rangle - f(x)\}.$$

The f^* is convex and lower-semicontinous as the supremum of of continuous affine functions. If f proper, f^* proper closed convex.

Example. Let $f(x) = \frac{1}{p}\|x\|^2, 1 < p < \infty,$

then $f^*(y) = \frac{1}{q}\|x\|_q^q, \frac{1}{p} + \frac{1}{p} = 1.$

We can define the *biconjugate* f^{**} as the conjugate of conjugate f^*.

$$f^{**}: X \to \mathrm{R} \cup \{+\infty\}$$

$$f^{**}(x) = sup_{x \in X^*}\langle x^*, x\rangle - f^*(x^*)$$

The f^{**} is the largest convex l.s.c function below f. It is easy to see from the definition and fenchel inequality that $f^{**} \leq f$.

$$f^{**}(x) \leq \langle x^*, x\rangle - f^*(x^*) \leq f(x)$$

Proposition 4.7.1. (Fenchel – Young Inequality). Let $f: X \to \mathbb{R} \cup \{+\infty\}$. For all $x \in X$ and $x^* \in X^*$, we have

$$f(x) + f^*(x^*) \geq \langle x^*, x\rangle.$$

Proof. Since x is not necessarily the maximizing point for $f(x^*)$=sup_x ..., we have $f(x^*) \geq \langle y, x\rangle - f(x) \Leftrightarrow f(x) + f(x^*) \geq \langle x^*, x\rangle$. ⊡

Note. The inequality holds $\Leftrightarrow x^* \in \partial f(x)$.

Proposition 4.7.2. When $f \leq g$, we have $f^* \geq g^*$. In particular,

$(sup_{i \in I} f(_i))^* \leq inf_{i \in I}(f_i^*)$ and $\left(inf_{i \in I}(f_i)\right)^* = sup_{i \in I}(f_i^*),\ \forall\ (f_i)_{i \in I}$ of functions on X with values in $\mathbb{R} \cup +\infty$ [1]. Proposition 4.7.2. is necessary to prove the next proposition, which help us to define primal – dual algorithms more quickly. In

particular we can replace the f by f^{**} if f is proper, convex and lower – semicontinuous.

Proposition 4.7.3. Let a function $f: X \to \mathbb{R} \cup \{+\infty\}$ which is proper. The function f is convex and lower-semicontinuous if, and only if, $f^{**} = f$.

Proof. [1] ($\Rightarrow$) Since f is convex and l.s.c., we can write f as the supremum of continuous and affine functions on X , we have $f = sup_{i\in I}(f_i)$. From previous proposition and $f \leq g$:

$$f^* \geq g^* \Rightarrow \mathrm{f}^{**} \leq \mathrm{g}^{**}.$$

Therefore, $f^{**} \geq sup_{i\in I}(f^{**}) = sup_{i\in I}(f_i) = f$, because $f_i^{**} = f_i$ is continuous and affine functions, and as we know $f^{**} \leq f \Rightarrow f^{**} = f$

($\Leftarrow$) Since $f^{**} = f$ is a supremum over the set of continuous affine functions. ⊡

An interesting consequence is the fact that, if f is convex, proper and l.s.c. then we have

$$f(x) + f^*(x^*) = \langle x^*, x\rangle \Leftrightarrow x^* \in \partial f(x).$$

By definition, we see that:

$$x \text{ realizes the } sup_{x\in X}\langle x^*, x\rangle - f(x) \Leftrightarrow x^* \in \partial f(x)$$

and we have

$$f(x) + f^*(x^*) = \langle x^*, x\rangle \Leftrightarrow f^{**}(x) = f(x) = \langle x^*, x\rangle - f^*(x) \Leftrightarrow x \in \partial f^*(x^*).$$

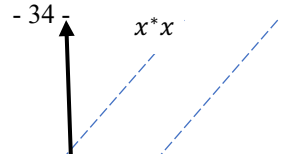
We can say that ∂f and ∂f^* are inverses,

$$x^* \in \partial f(x) \Leftrightarrow x \in \partial f^*(x^*).\boxdot$$

In this point is good to refer that conjugates functions do not give us anything new itself, it helps to derive the dual problem more quickly.

Geometry of Conjugates

We assume a function $f: \mathbb{R} \to \mathbb{R}$ the interpretation of conjugate $f^*(x^*)$ is: for the function $f(x)$, given a x^*, we assume a line $h(x) = xx^*$[11]. We want to find a value on the $x - axis$ such that, the value x maximizes the difference between the line $h(x)$



and function $f(x)$. Let $\hat{x}$ be the optimal value, we define a parallel line g to h, which is passing through the point $(y, f(y))$. The intercept of g and $y - axis$ is the $-f^*(x^*)$.

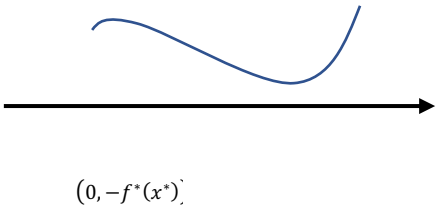


Figure 7 Geometry of Conjugate

4.8. Fenchel – Rockafellar Duality

This notion is very useful. It helps us to transform convex problems into others with better properties, which are easier to handle them. In this subsection we assume that X, Y are normed spaces and $K \in X^*$, is a linear and bounded operation.

Let $f: X \to \mathbb{R} \cup \{+\infty\}$, $g: X \to \mathbb{R} \cup \{+\infty\}$ be proper, convex and lower-semicontious. We define the *primal problem* (PP) as

$$inf_{x \in X} f(Kx) + g(x).$$

We prove (Proposition 4.7.3) for f proper, convex and l.s.c. that $f^{**} = f$. We replace the f by f^{**} and rewrite the primal problem as

$$inf_{x \in X} f(Kx) + g(x) = inf_{x \in X} sup_{y \in Y} \langle y, Kx \rangle - f^*(y) + g(x).$$

Theorem 4.8.1. Let X be a convex subset of a linear topological space, Y be a compact convex subset of a linear topological space, and $f: X \times Y \to \mathbb{R}$ an upper semicontinuous on X and lower semicontinuous on Y. Suppose that f is quasiconcave on X and quasiconvex on Y. Then we have,

$$min_Y sup_X f = sup_X min_Y f. \boxdot$$

From the above theorem we can swap min and sup and we have [3],

$$inf_x f(Kx) + g(x) = inf_x sup_y \langle y, Kx \rangle - f^*(y) + g(x)$$

$$= sup_y inf_x \langle y, Kx \rangle - f^*(y) + g(x)$$

$$= sup_y - f^*(y) - g^*(-K^*y).$$

The last formula is known as *dual problem* (DP). Therefore, the primal is equal to dual and the $sup_y inf_x \langle y, Kx \rangle - f^*(y) + g(x)$ problem, is the *primal-dual* problem. The y^* is the solution of dual problem and x^* is the solution of the initial primal problem. The solution (x^*, y^*) is a saddle point of the primal-dual problem. We define the *Lagrangian* as the $\mathcal{L}(x, y) \coloneqq \langle y, Kx \rangle - f^*(y) + g(x)$. [3] The *saddle point* of the primal-dual problem is any pair $(x, y) \in X \times Y$, such that

$$\mathcal{L}(x^*, y) \leq \mathcal{L}(x^*, y^*) \leq \mathcal{L}(x, y^*).$$

The *primal dual gap* is defined as

$$\mathcal{g}(x, y) \coloneqq f(Kx) + g(x) + f^*(y) + g^*(-K^*y)$$

$$= sup_{(x', y') \in (X \times Y)} \mathcal{L}(x, y') - \mathcal{L}(x', y).$$

If (x^*, y^*) is a saddle point the primal dual gap is zero. The *optimally conditions* are

$$\begin{cases} 0 \in \partial g(x^*) + K^* y^* \\ 0 \in \partial f^*(y^*) - K x^*. \end{cases}$$

5 ALGORITHMS

5.Introduction

In this chapter we discuss the basic algorithms for solving convex optimization problems. These algorithms are iterative procedures. We will discuss also their convergence. First, we analyze the gradient method, which minimize function, where are differentiable. Then, we will see how to handle functions non smooth with the proximal point method and combining the two methods we have the proximal gradient method, which handles decomposable function with smooth and non-smooth functions. Finally, we study the primal dual algorithm.

5.1. Iterative Procedures

An *iterative algorithm* on X [1] is a procedure by which, starting from an initial point $x_0 \in X$, and using a family (T_n) of functions from X to X,

$$x_{n+1} = T_n(x_n)\ \forall n \geq 0,$$

we construct a sequence $x_n \in X$.

These procedures help us to find minimizers of a function f. The idea is, each time, to find a point x_{n+1} where $f(x_{n+1}) < f(x_n)$, for this reason we are moving in a specific direction and we construct a sequence, which minimize the function f.

In this point, let discuss the issue of convergence of the sequences. [1] We know that on a Banach space all sequences are Cauchy and therefore we have convergence. Hilbert space is a Banach space and this is useful to prove weak convergence of a sequence in Hilbert spaces.

Lemma 5.1.(Opial's Lemma) [1] Let $S \subset \mathbb{H}, argmin(f) \neq \emptyset$, and $(z_n) \in \mathbb{H}$. We assume:

(a) For each $u \in argmin(f)$ there exists $lim_{n\to\infty}\|x_n - u\|$

(b) Every weak limit point of (z_n) belongs to $argmin(f)$.

Then (z_n) converges weakly as $n \to \infty$ to some $\hat{u} \in argmin(f)$

5.2. Gradient Method

In this subsection we describe the gradient method. This method helps us to minimize convex and differentiable functions. This method is a first-order method. The idea is

that the function f decreases fastest if one goes from a point $x \in dom(f)$ in the direction of the $-\nabla f(x)$. This implies that for the iterative sequence

$$x_{n+1} = x_n - \lambda_n \nabla f(x_n), n \geq 0,$$

we have that $f(x_n) \geq f(x_{n+1})$. We want to move against the gradient of f, toward the minimum. We set an initial x_0 and we construct a sequence (x_n) such that

$$x_{n+1} = x_n - \lambda_n \nabla f(x_n), n \geq 0.$$

As we say the $f(x_n)$ is monotonic sequence, the question is what is holds with convergence. Under curtain assumptions like f convex, ∇f Lipschitz continuous and the step sizes λ_n particularly chosen we assure the convergence.

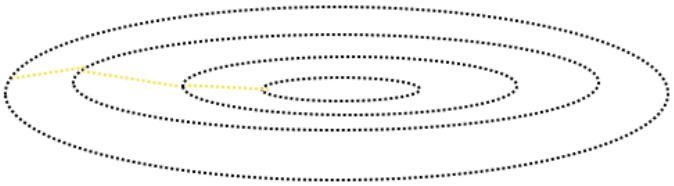


Figure 8 Gradient Method

Let $f: \mathbb{H} \to \mathbb{R}$ be continuously differentiable function with Lipschitz-continuous ∇f. Let the ordinary differential equation:

(ODE) $$\begin{cases} x(0) = x_0 \\ -\dot{x(t)} = \nabla f\big(x(t)\big), t > 0 \end{cases}.$$

By the Cauchy-Picard Theorem, for each $x_0 \in \mathbb{H}$, the (ODE) has a unique solution, there is a unique continuously differentiable function $x: [0, +\infty) \to \mathbb{H}$ such that $x(0) = x_0$ and $-\dot{x(t)} = \nabla f\big(x(t)\big)$ for all $t > 0$. [1] The stationary points of (ODE) are the zeroes of gradient of f. [7] The (ODE) solves the problem of minimizing f in the sense that for every trajectory $x(t)$, we have $f\big(x(t)\big) \to \hat{z}$. The function f decreases along the solutions [1]. Decreases strictly into a critical point, for more details see [1].

From the fact that f is nonincreasing, we have that

$$lim_{t\to\infty} f\big(x(t)\big) = inf(f).$$

If assume also that we have at least one minimizer of f and we take $\hat{z} \in argmin(f)$, $lim_{t\to\infty} \|x(t) - \hat{z}\|$ exists. From proposition f is weakly lower semicontinuous because is convex and continuous. And every weak limit of $x(t)$ must minimize f, as

$t \to \infty$. Finally, from Opial's Lemma $x(t) \to \hat{z} \in S$, as $t \to \infty$ (weakly), see [1] for more details. ⊡

We discretize (ODE) [1] with finite differences, and the reason is to approximate $\dot{x}(t)$

- Let (λ_n) be positive parameters, called *step sizes*.
- Set $\sigma_n = \sum_{k=1}^{n} \lambda_k$ and
- The partition of $[0, +\infty) = \bigcup_{n=1}^{\infty} \sigma_n$, where $\lambda_i = \sigma_i - \sigma_{i-1}, i = 0, \dots, n, \dots$

We assume $t \to \infty, \sigma_n \to \infty, n \to \infty \Leftrightarrow \lambda_n \in \ell^1$. Now we approximate $\dot{x}(t)$ by

$$\frac{x_n - x_{n-1}}{\lambda_n}.$$

If we approximate the term $\nabla f\big(x(t)\big)$ by $\nabla f(x_{n-1})$ we have from (ODE) that

$$-\frac{x_n - x_{n-1}}{\lambda_n} = \nabla f(x_{n-1}) \Leftrightarrow x_n = x_{n-1} - \lambda_n \nabla f(x_{n-1}).$$

This method, with this update step is called *gradient method* and is applied on differentiable functions.

With the same logic we can approximate the term $\nabla f\big(x(t)\big)$ by $\nabla f(x_n)$ and we have,

$$-\frac{x_n - x_{n-1}}{\lambda_n} = \nabla f(x_n) \Leftrightarrow x_{n-1} = x_n + \lambda_n \nabla f(x_n).$$

This method is known as *proximal method,* and is a generalization of gradient to non-smooth functions. We shall discuss this method on the next subsection. ⊡

Let $f: \mathbb{H} \to \mathbb{R}$ be convex, with ∇f Lipschitz continuous with constant L. The *(pure) gradient method*, is starting from an initial point $x_0 \in \mathbb{H}$ and we apply the iteration step

$$x_{n+1} = x_n - \lambda_n \nabla f(x_n), for\ n \in \mathbb{N}$$

With condition for the step sizes be

$$sup_{n \in \mathbb{N}} \lambda_n < \frac{2}{L}.$$

Therefore, the idea of this iterative algorithm is:

Algorithm 1 Gradient Method (G)

Choose $x_0 \in \mathbb{H}$

for all $n \geq 0$ **do**

$$x_{n+1} = x_n - \lambda_n \nabla f(x_n)$$

end for

A proximal point of Gradient method (G) [9]

By Taylor expansion, in each iteration we can consider the expression,

$$f(x_{n+1}) \approx f(x_n) + \langle \nabla f(x_n), x_{n+1} - x_n \rangle + \frac{1}{\lambda_n} \|x_{n+1} - x_n\|^2,$$

Where the term $f(x_n) + \langle \nabla f(x_n), x_{n+1} - x_n \rangle$ is a linear approximation and the term $\frac{1}{\lambda_n} \|x_{n+1} - x_n\|^2$ is the proximity term (it is replacing the hessian matrix), therefore we can express the $x_{n+1} = argmin_x f(x_n) + \langle \nabla f(x_n), x - x_n \rangle + \frac{1}{\lambda_n} \|x - x_n\|^2$, λ_n are step sizes. The geometrical interpretation of this expression is

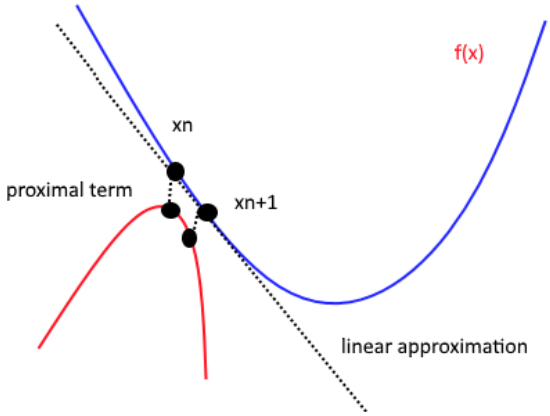


Figure 9 A proximal point of Gradient method

If λ_n is small, x_{n+1} tends to stay close to x_n.

Convergence of Gradient method (G).

The convergence of gradient method is succeeding under the next assumptions. We assume f be convex, differentiable, ∇f be Lipschitz continuous and with specific choice of step sizes we have the next theorem [1].

Theorem 5.2.1. [1] Let (x_n) satisfy (G), where f is convex, $S \neq \emptyset$, $\lambda_n \notin \ell^1$ and $sup_{n \in \mathbb{N}} \lambda_n < \frac{2}{L}$. Then (x_n) *converges weakly* as $n \to \infty$ to point in S.

Note that we have strongly convergent $\Leftrightarrow$ f is strongly convex or f is even or $int\big(argmin(f)\big) \neq \emptyset$.

We know and the rate of convergence from the next theorem

Theorem 5.2.2. [3] Let f convex and gradient of Lipchitz continuous with constant L and $k < n$. Gradient algorithm with fix step size (step size doesn't change after each iteration) $\lambda < \frac{1}{L}$ satisfies

$$f(x_k) - f(x^*) \leq \frac{\|x^0 - x^*\|^2}{2k\lambda},$$

where x^* is any minimizer of f. If in addition f is strongly convex with parameter $\mu > 0$, we have

$$f(x_k) - f(x^*) \leq \omega^{\kappa} \frac{L}{2} \|x_0 - x^*\|^2.$$

Therefore, we have,

- If f convex the convergence rate is $O\left(\frac{1}{k}\right)$
- If f is μ −strongly convex the convergence rate is $O(\omega^k)$

Details about proof is on [3].

It is obvious that if f is strongly convex the algorithm is very fast.

The gradient method is for C^1-smooth and unconstrained problems. The gradient method is a simple idea and under special assumptions is fast but if the function isn't strongly convex is slow and cannot handle non-smooth functions.

5.3. Proximal Point Algorithm

Let $f: \mathbb{H} \to \mathbb{R} \cup \{+\infty\}$ is proper, lower-semicontinous convex function. Let (λ_n) is positive numbers. They are called *step sizes*.

The idea is to minimize the Moreau – Yosida Regularization $f_{(\tau_n, x_n)}$ of f, which is proper, lower-semicontinuous and strongly convex function and has unique minimizer. We construct a sequence as the next one:

$$.x_{n+1} = argmin\{f(z) + \frac{1}{2\lambda_n}\|z - x_n\|^2\}.$$

[1] By the Moreau – Rockafellar Theorem and because x_{n+1} is the minimum we have,

$$0 \in \partial f_{(\lambda_n, x_n)}(x_{n+1}) = \partial f(x_{n+1}) + \frac{x_{n+1} - x_n}{\lambda_n}$$

$$\Leftrightarrow -\frac{x_{n+1} - x_n}{\lambda_n} \in \partial f(x_{n+1})$$

$$\Leftrightarrow x_{n+1} = (I + \lambda_n \partial f)^{(-1)}(x_n).$$

This sequence (x_n) is called *proximal sequence*. The stationary points of a proximal sequence are the minimizers of the objective function [1] since,

$$x_{n+1} = x_n \Leftrightarrow 0 \in \partial f(x_{n+1}).$$

At this point it is good to mention that the proximal point algorithm can be interpreted as discretization of the differential inclusion [1]

$$-\dot{x}(t) \in \partial f\big(x(t)\big) \quad t > 0.$$

From the definition of proximal point algorithm, [1] we have

$$f(x_{n+1}) + \frac{1}{2\tau_n}\|x_{n+1} - x_n\|^2 \leq f(x_n) \ \forall n.$$

Therefore, the sequence $\big(f(x_n)\big)$ is nonincreasing.

Recall the notion of proximity operator from subsection 4.6.

$$prox_{\lambda f}(\mathrm{x}) = argmin_{x\in\mathbb{H}} f(z) + \frac{1}{2\lambda}\|x - z\|^2.$$

The update step of proximal point algorithm (**PPA**) is

$$x^{k+1} = prox_{\lambda f}(x^k).$$

The proximal method is for smooth and non-smooth problems, constrained and unconstrained problems. The $prox_{\lambda f}$ is a convex optimization problem that uses the proximal operator of the objective functions.[7] The **PPA** minimizes a convex function f by repeatedly applying the $prox_{\lambda f}$ to some initial x_0.

Algorithm 2. Proximal Point Algorithm (PPA)

---

choose $x_0 \in \mathbb{H}$

for $k = 0,1, \ldots$

$$x^{k+1} = prox_{\lambda f}(x^k).$$

end for.

From the next proposition which is in [1] we have that the direction of x_n is towards to the set $argmin(f)$.

Proposition 5.3.1. Let (x_n) be a proximal sequence. If $x_{n+1} \neq x_n$, then

$$\langle x^{n+1} - x^n, x^n - x^{n-1}\rangle > 0.$$

Additionally, if we have, $\hat{x} \in argmin(f)$ then

$$\langle x^{n+1} - x^n, \hat{x} - x^n\rangle > 0. \ \square$$

The proximity operator

The notion of the proximal operator,

$$prox_{\lambda f}(\mathrm{x}) = argmin_{x\in\mathbb{H}} f(z) + \frac{1}{2\lambda}\|x - z\|^2 ,$$

illustrated in figure 10. The black lines are the level curves of the function and the bold black is the boundary [7]. We calculate the $prox_{\lambda f}$ to the blue points and then they have moved to red.

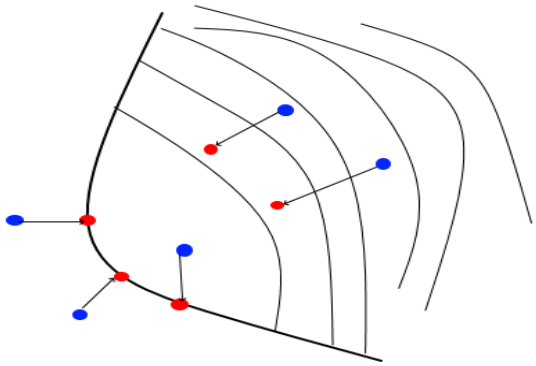


Figure 10 Interpretation of proximal operator

The step size (parameter) λ controls how fast we move towards the minimum. Large values provide big steps to the minimum and small values small.

After all, it is obvious that $prox_{\lambda f}(v)$ is a point between the minimum of f and a point $v \in dom(f)$.

Example 5.3.2. Let δ_C the indicator function. The proximal operator of the indicator function is

$$prox_{\delta_C}(x) = argmin_y \left(\delta_C(y) + \frac{1}{2}\|y - x\|^2\right)$$

$$= argmin_{y\in C}\frac{1}{2}\|y - x\|^2$$

$$=: proj_C(x).$$

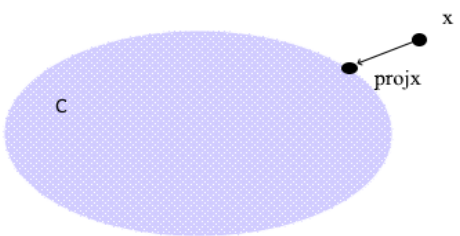


Figure 11 Proximal Operator of Projection

In some sense, we can say that proximal iteration generalizes the notion of projection, when the function $f(x)$ is not the indicator but a lower-semicontinuous and convex function.

Calculation of proximal operator.

Let $f: \mathbb{H} \to \mathbb{R} \cup \{+\infty\}$, for $x \in \mathbb{H}$ and $\lambda > 0$ we can find $y \in \mathbb{H}$ such that

$$y \in argminf(v) + \frac{1}{2\lambda}\|v - x\|^2 : v \in \mathbb{H} \Leftrightarrow x - y \in \lambda\, \partial f(y)$$

Example 5.3.3. [16] ℓ_1-norm

Let $f(x) = \|x\|_1$. Then $prox_{\lambda f}(v) = argmin_{x \in X}(\|x\|_1 + \frac{1}{2\lambda}\|x - y\|^2)$.

We have, from Fermat's rule that

$0 \in \partial f(v^*) + \frac{1}{\lambda}(v^* - v) \Leftrightarrow v - v^* \in \lambda\, \partial f(v^*)$ (by the subgradient condition).

Recall from subgradient of ℓ_1-norm $\partial f(x) = \partial|x_1| \times \ldots \times \partial|x_n|$ this implies that

$$\left(prox_{\lambda f}(v)\right)_i = \begin{cases} v_i - \lambda, v_i \geq \lambda \\ 0, \quad |v_i| \leq \lambda \\ v_i + \lambda, \leq -\lambda. \end{cases}$$

Finally, the $prox_{\lambda f}(v) = shrink(v, \lambda)_i = max(|v_i| - \lambda, 0)\frac{v_i}{|v_i|}$. This operator is known as the *soft thresholding* operator.

Example 5.3.4. Let $f(x) = \|x\|_2$, then $prox_{\lambda f}(x) = max(\|x\|_2 - \lambda, 0)\frac{x}{\|x\|_2}$. This sometimes is called *block soft thresholding* operator.

Convergence of Proximal Point Algorithm (PPA)

Theorem 5.3.5. [2] Let $f: \mathbb{H} \to \mathbb{R} \cup +\infty$ and $argmin(f) \neq \emptyset$, let (λ_n) be the sequence of step sizes such that $\sum_{n \in \mathbb{N}} \lambda_n = +\infty$, and let $x_0 \in \mathbb{H}$. Let the proximal iteration

$$(\forall n \in \mathbb{N})\ x_{n+1} = prox_{\lambda_n f} x_n \quad (5.8)$$

Then the following statements hold:

(a) (x_n) is a minimizing sequence of f, $f(x) \downarrow inf\ f(\mathbb{H})$.

(b) (x_n) converges weakly to a point $\hat{x} \in argmin(f)$.

Proof. (a)Let $y \in S$. It follows from definition of x_n (5.8) and from the optimality condition

$$x_n - x_{n+1} \in \lambda_n \partial f(x_{n+1}).$$

From (16.1) we have,

$$\frac{1}{\lambda_n}\langle y - x_{n+1}, x_n - x_{n+1}\rangle \leq f(y) - f(x_{n+1}) \text{ (5.9)}$$

And

$$0 \leq \frac{1}{\lambda_n}\langle x_n - x_{n+1}, x_n - x_{n+1}\rangle \leq f(x_n) - f(x_{n+1}).$$

From (5.9) for every $n \in \mathbb{N}$, we have

$$\begin{aligned}\|x_{n+1} - y\|^2 &\leq \|x_n - y\|^2 + \langle y - x_{n+1}, x_n - x_{n+1}\rangle + \|x_{n+1} - x_n\|^2 \\ &= \|x_n - y\|^2 - \|x_{n+1} - x_n\|^2 + \langle x_{n+1} - y, x_{n+1} - x_n\rangle \\ &\leq \|x_n - y\|^2 - 2\lambda_n\big(f(x_{n+1}) - inff(\mathbb{H})\big).\end{aligned}$$

This implies that x_n is Fejer-Monotone with respect to $argmin(f)$ and

$$\sum_{n\epsilon\mathbb{N}} 2\lambda_n\big(f(x_{n+1}) - inff(\mathbb{H})\big) < +\infty.$$

Since, $\sum_{n\in\mathbb{N}} \lambda_n = +\infty$ we have $f(x_n) \downarrow inff(\mathbb{H})$.

(.b) Let $\hat{x}$ be a weak point of x_n. It follows from the next proposition.

Proposition 5.3.6. Let f be proper, l.s.c. quasiconvex function and let (x_n) be a minimizing sequence of f that converges weakly to $\hat{x} \in H$. Then $f(x) = inf(H)$.From previous proposition and theorem 3.4.3. the proof is complete.

5.4. Proximal Gradient Method

As we say the proximal operator can handle non smooth function. Consider the problem

$$minf(x) + g(x) \text{ (5.10)}$$

Where $f: \mathbb{H} \to \mathbb{R} \cup \{+\infty\}$, $g: \mathbb{H} \to \mathbb{R} \cup \{+\infty\}$ are closed, convex, proper function. Let f be differentiable but g be non smooth.

The proximal gradient method is

$$x_{k+1} \coloneqq prox_{\lambda_k g}\Big(x_k - \lambda_k \nabla f(x^k)\Big),$$

Where, $k < n$ the number if iterations and λ_k is a step size.

Algorithm 3. Proximal Gradient Method

---

choose $x_0 \in \mathbb{H}$

for $k = 0,1 \ldots$

$$x_{k+1} \coloneqq prox_{\lambda_k g}\Big(x_k - \lambda_k \nabla f(x^k)\Big),$$

end for.

The update step is like searching fixed point of proximal operator. In the sense that if $\hat{x}$ is a solution of (5.10), [7] by the optimality condition, $\hat{x}$ must satisfy

$$0 \in \nabla f(\hat{x}) + \partial g(\hat{x})$$
$$\Leftrightarrow 0 \in \nabla f(\hat{x}) + \partial g(\hat{x}) - \hat{x} + \hat{x}$$
$$\Leftrightarrow (I + \lambda\, \partial g)(\hat{x}) \ni (I - \lambda \nabla f)(\hat{x})$$
$$\Leftrightarrow \hat{x} = (I + \lambda\, \partial g)^{-1}(I - \lambda \nabla f)(\hat{x})$$
$$\Leftrightarrow \hat{x} = prox_{\lambda g}\big(\hat{x} - \lambda \nabla f(\hat{x})\big)$$

The last equality says that $\hat{x}$ minimizes the problem $(5.10) \Leftrightarrow$ is a fixed point of the *forward – backward operator* $(I + \lambda \partial g)^{-1}(I - \lambda \nabla f)$.

Convergence of Proximal Gradient Method

Theorem 5.4.1. We assume that ∇f is Lipschitz continuous with constant $L > 0$ and the step sizes are $\lambda \leq \frac{1}{L}$, then we have

$$f(x_k) - f^* \leq \frac{\|x_0 - x^*\|^2}{2\lambda k} \cdot \boxdot$$

This theorem implies that the proximal gradient has convergence rate $O\left(\frac{1}{k}\right)$. The reason why we have the condition $\lambda \in \left(0, \frac{1}{L}\right]$ implies that the operator $(I + \lambda \partial g)^{-1}(I - \lambda \nabla f)$ is averaged [7] and thus that the iteration convergence to a fixed point, with the assumption that exists one. Is a consequence from the next theorem.

Theorem 5.4.2. (The Baillon-Haddad Theorem) Let $f: H \to \mathbb{R}$ be convex and GD on H, and its gradient operator $\nabla f(x)$ nonexpansive. Then f is Fréchet differentiable and ∇f is firmly nonexpansive [6].

A very important question about selection of step sizes when we don't know the Lipschitz constant or is complicated to evaluated. [7] We can find step size by a line search. We take a parameter $b \in (0,1)$ and at each iteration we change the step sizes as

$$\lambda = b \cdot \lambda.$$

Recall that for function f from Taylor expansion we have an upper bound. In particular, the function f is bounded from above from the function

$$\hat{f}_\lambda(x,y) = f(y) + \langle f(y), x-y\rangle + \frac{1}{2\lambda}\|x-y\|^2, \text{ with } \lambda > 0.$$

We can apply the proximal method as the pseudocode [7]:

Given $x_k, \lambda_{\kappa+1}, b \in (0,1)$

Let $\lambda \coloneqq \lambda_{k+1}$

Repeat

1. Let $z \coloneqq prox_{\lambda g}\big(x_k - \lambda \nabla f(x_k)\big)$
2. Break if $f(z) \leq \hat{f}_\lambda(z, x_k)$
3. Update $\lambda \coloneqq b\lambda$.

Return $\lambda_k = \lambda$, $x_{k+1} \coloneqq z$

5.5. Accelerated proximal gradient.

A method to make the proximal gradient method faster is to add an extrapolation step. Then the algorithm is

$$y_{k+1} \coloneqq x_k + \omega_k(x_k - x_{k-1})$$
$$x_{k+1} \coloneqq prox_{\lambda_k g}\big(y_{k+1} - \lambda_k \nabla f(y_{k+1})\big)$$

Where $\omega_k \in [0,1)$ and is called extrapolation parameter and λ_k is the step sizes. Obtain that for $k = 1$ we have the usual proximal gradient method, but for next iterations the y_{k+1} has some information from previous iteration. Some usual choices of extrapolation parameter are $\omega_k = \frac{k-1}{k+2}$ or $\omega_k = \frac{k}{k+3}$. The convergence rate of accelerated proximal gradient algoritm is $O\left(\frac{1}{k^2}\right)$.

5.6. Primal dual Algorithm.

Recall from chapter 4 the Fenchel-Rockafellar duality. We write the problem

$$inf_{x \in X} f(Kx) + g(x),$$

where f, g are convex, and $K: X \to Y$ is a bounded, linear operator, as the *primal – dual problem*

$$min_y inf_x \langle y, Kx\rangle - f^*(y) + g(x).$$

A good reason why we need this formula is when f is non – smooth function but we can take proximal operators of f^* and g easily.

The idea is to swing a descent step for primal variable x and an ascent step for dual variable y.

Algorithm 4. Primal-Dual

Input: initial point (x_0, y_0), steps $\sigma > 0, \tau > 0$, so that $\sigma\tau L^2 < 1$, where $L = \|K\|$, and $\theta \in [0,1]$.

for all $k \geq 0$ **do**

$find(x_{k+1}, y_{k+1})$ by solving

$$y_{k+1} = prox_{f^*}(y_k + \sigma K\bar{x}_k) \text{ (dual proximal)}$$

$$x_{k+1} = prox_g(x_k - \tau K^* y_{k+1}) \text{ (primal proximal)}$$

$$\bar{x}_{k+1} = x_{k+1} + \theta(x_{k+1} - x_k) \ (extrapolation)$$

end for

The extrapolation step helps us to have convergence. The convergence rate it depends on the type of problem:

- If the problem is non smooth: $O\left(\frac{1}{N}\right)$
- Sum of a smooth and non-smooth: $O\left(\frac{1}{N^2}\right)$
- If the problem is smooth: $O(\omega^N), \omega < 1$

6 Minimization of Lasso function

6.Introduction

In this chapter we will calculate the proximal operators for Lasso problem. We will simulate data in MATLAB and run the algorithms of proximal gradient and accelerated proximal. Then we will compare the time and the iterations each method needs. Finally, we calculate the dual of LASSO.

6.1. LASSO

The *Lasso problem* is

$$minimize \frac{1}{2}\|Ax - b\|_2^2 + \gamma\|x\|_1,$$

where $x \in \mathbb{R}^n$, where $A \in \mathbb{R}^{m\times n}$, and $\gamma > 0$.

We will treat this problem in the Hilbert space $\mathbb{R}^n$ endowed with the $\ell_2 -$ norm.

Proposition 6.1.1. The objective of Lasso is convex.

Proof. Let $f(x) = \frac{1}{2}\|Ax - b\|_2^2 + \gamma\|x\|_1$. We can write $f(x)$ as $f(x) = h(x) + g(x)$, where $h(x) = \frac{1}{2}\|Ax - b\|_2^2$ and $g(x) = \gamma\|x\|_1$. Note that $dom(h) = \mathbb{R}^n$ and $dom(g) = \mathbb{R}^n$ and both domains are convex sets.

Convexity of $h(x)$. The Hessian of $h(x)$ is $\nabla^2 f(x) = A^T A$. The Hessian is positive semidefine, since for any $x \in \mathbb{R}^n$ we have $x^T A^T Ax = \|Ax\|_2^2 \geq 0$. Hence, the function $h(x)$ is convex.

Convexity of $g(x)$. For any x_1, x_2 and any $\theta \in (0,1)$, let $x = \theta x_1 + (1-\theta)x_2$. Then

$$\begin{aligned} g(x) &= \gamma\|\theta x_1 + (1-\theta)x_2\| \\ &\leq \gamma\|\theta x_1\| + \gamma\|(1-\theta)x_2\| \\ &= \gamma\theta\|x_1\| + \gamma(1-\theta)\|x_2\| \\ &= \theta g(x_1) + (1-\theta)g(x_2) \end{aligned}$$

Hence $g(x)$ is convex. As we know the sum of two convex function is a convex function, therefore, $f(x) = h(x) + g(x)$ is also convex. ⊡

In general, the Lasso problem can be interpreted as finding a sparse solution to a linear regression model or to a least squares problem, where this implies a variable selection method.

6.2. Proximal gradient method

For Lasso problem, let $f(x) = \frac{1}{2}\|Ax - b\|_2^2$ and $g(x) = \gamma\|x\|_1$. The function $f(x)$ is differentiable but function $g(x)$ is non smooth. The gradient of f is:

$$\nabla f(x) = A^T(Ax - b)\ .$$

Recall now that the proximal of ℓ_1 norm is the soft thresholding operator is:

$$\left[S_\gamma(x)\right]_i = \left(prox_{\gamma g}(x)\right)_i = \begin{cases} x_i - \gamma, x_i \geq \gamma \\ 0, \quad |x_i| \leq \gamma \\ x_i + \gamma, x_i \leq -\gamma. \end{cases}$$

Hence the proximal operator for function $g(x)$ is:

$$prox_{\gamma g}(x) = S_\gamma(x)$$

By definition of the proximal gradient, the iteration is given from the formula:

$$x_{k+1} = prox_{\lambda_k g}\left(x_k - \lambda_k \nabla f(x_k)\right)$$

Therefore, the proximal gradient update is

$$x_{(n+1)} = S_{\gamma\lambda}\left(x_n + \lambda A^T(b - Ax_n)\right).$$

This algorithm is called *iterative-soft thresholding algorithm* **(ISTA)**. The accelerated version of ISTA is called **FISTA**. [12]

In the next table we compare the algorithms ISTA and FISTA, for simulated data from normal distribution $N(0,1)$ and regularization parameter $\gamma = 0.1\gamma_{max}$, $\gamma_{max} = \|A^T b\|_\infty$. [7]

Method	*Iterations*	*Time (s)*	$\boldsymbol{p}^*$
ISTA	143	8.3344	21.188
FISTA	108	7.3175	21.220

6.3. Primal-Dual Problem

Recall the Lasso problem

$$min_{x\in\mathbb{R}^p}\ \frac{1}{2}\|b - Ax\|_2^2 + \gamma\|x\|_1.$$

By theory of primal dual we add an auxiliary variable $y = Ax$, and the Lasso problem is equivalent to

$$min_{y,x}\ \frac{1}{2}\|b - y\|_2^2 + \gamma\|x\|_1 \text{ subject to } Ax = y.$$

The Lagrangian is $L(y, x, u) = \frac{1}{2}\|b - y\|_2^2 + \gamma\|x\|_1 + \langle u, (y - Xx)\rangle$, where u is the dual variable and x, y is primal variables. Now we want to minimize the $L(y, x, u)$.

$$max_{\lambda} min_{x,} \gamma \|x\| + \langle \lambda, Ax \rangle - h^*(\lambda)$$

Where $h^*(\lambda) = h(\lambda)$, $h^*(\lambda) = \frac{1}{2}\|b - A\lambda\|_2^2$.

The update steps are:

$$y_{k+1} = prox_{h*}(y_k + \sigma A x_k)$$

$$x_{k+1} = prox_g(x_k - \tau A^* y_{k+1})$$

$$\bar{x} = x_{k+1} + \theta(x_{k+1} - x_k)$$

Recall, in this point, that the proximal operator of h^* is the *block soft thresholding operator*.

APPENDICES

Code for Matlab [17]

https://web.stanford.edu/~boyd/papers/prox_algs/lasso.html#6

```
function p = objective(A, b, gamma, x, z)
    p = 0.5*sum((A*x - b).^2) + gamma*norm(z,1);
end

function s = prox_l1(v, lam)

    s = max(0, v - lam) - max(0, -v - lam);
end
m = 500;        % number of examples
n = 2500;       % number of features

%x1 = sprandn(n,1,0.05);

%A = randn(m,n);

%A = A*spdiags(1./sqrt(sum(A.^2))',0,n,n); % normalize
columns
%v = sqrt(0.001)*randn(m,1);
%b = A*x1 + v;

myx=x1;
save myfile.mat
myA=A;
save myfile.mat
myv=v;
save myfile.mat
myb=b;
save myfile.mat
load myfile.mat
myx;
load myfile.mat
myA;
load myfile.mat
myv;
load myfile.mat
myb;

x0=myx;
A=myA;
v=myv;
b=myb;

gamma_max = norm(A'*b,'inf');
```

```
gamma = 0.1*gamma_max;

% cached computations for all methods
AtA = A'*A;
Atb = A'*b;

MAX_ITER = 300; % to k sto for tha mas deiksei poses
xreiastikan
ABSTOL   = 1e-4;
RELTOL   = 1e-2;

f = @(u) 0.5*sum((A*u-b).^2);

%ISTA
lambda = 1;
beta = 0.5;

tic;

x = zeros(n,1);
xprev = x;


for k = 1:MAX_ITER
    while 1
        grad_x = AtA*x - Atb;
        p1=x - lambda*grad_x;
        p2=lambda*gamma;
        z = prox_l1(p1, p2);
        if f(z) <= f(x) + grad_x'*(z - x) +
(1/(2*lambda))*sum((z - x).^2)
            break;
        end
        lambda = beta*lambda;
    end
    xprev = x;
    x = z;

    h.prox_optval(k) = objective(A, b, gamma, x, x);
    if (k > 1 )&& abs(h.prox_optval(k) - h.prox_optval(k-
1)) < ABSTOL
        break;
    end
end

h.x_prox = x;
h.p_prox = h.prox_optval(end);
h.prox_grad_toc = toc;
h.p_prox
h.prox_grad_toc
k
```

```
%FISTA
lambda = 1;

tic;

x = zeros(n,1);
xprev = x;
for l = 1:MAX_ITER
    y = x + (l/(l+3))*(x - xprev);
    while 1
        grad_y = AtA*y - Atb;
        p3=y - lambda*grad_y;
        p4=lambda*gamma;
        z = prox_l1(p3, p4);
        if f(z) <= f(y) + grad_y'*(z - y) +
(1/(2*lambda))*(sum(z - y).^2)
            break;
        end
        lambda = beta*lambda;
    end
    xprev = x;
    x = z;

   h.fast_optval(l) = objective(A, b, gamma, x, x);
   if (l > 1) && abs(h.fast_optval(l) - h.fast_optval(l-1)) <
ABSTOL
       break;
   end
end

h.x_fast = x;
h.p_fast = h.fast_optval(end);
h.fast_toc = toc;
h.fast_toc %ctime to run
h.p_fast %optimal vaalue
l %iterations
```

References

[1] **Juan Peypouquet (2015)** *Convex Optimization in Normed Spaces, Theory, Methods and Examples.* Springer

[2] **Heinz H. Bauschke, Patrick L. Combettes (2010)** *Covnex Analysis and Monotone Operator Theory in Hilbert Spaces.* Springer

[3] **Antonin Chambolle, Thomas Poch (2016)** *An introduction to continuous optimization for imaging.* HAL archives-ouvertes

[4] **Stephen Boyd, Lieven Vandenberghe (2004)** *Convex Optimization.* Cambridge university press.

[5] **Ernest K. Ryu, Stephen Boyd**. *A primer on monotone operator methods survey.* Appl. Comput. Math., V.15, N.1, 2016, pp.3-34

[6] **Charles L.Byrne (November 24, 2014)** *On a Generalized Baillon-Haddad Theorem for Convex Functions on Hilbert Space.*

[7] **Neal Parikh, Stephen Boyd (2013)** *Proximal Algorithms.* Vol. 1, No. 3(2013) 123-231

[8] **Yuxin Chen (2018)** *Dual and primal-dual methods.* Princeton University. Lecture notes. http://www.princeton.edu/~yc5/ele522_optimization/lectures/dual_method.pdf

[9] **Yuxin Chen (2017)** *Lasso: Algorithms and Extensions*. Princeton University. Lecturenotes. http://www.princeton.edu/~yc5/ele538b_sparsity/lectures/lasso_algorithm_extension.pdf

[10] **Wotao Yin (2016)** *Convergence of Fixed-Point Iterations.* UCLA Math. Lecture notes.

https://www.math.ucla.edu/~wotaoyin/summer2016/5_fixed_point_convergence.pdf

[11] Ryan Tibshirani (2015) *10-725/36-725 Convex Optimization.* Lecture notes.

[12] **Ryan Tibshirani.** *10-725/36-725 Proximal Gradient Descent and Acceleration.* Lecture notes. http://www.stat.cmu.edu/~ryantibs/convexopt-F16/lectures/prox-grad.pdf

[13] **R. Tyrrell Rockafellar (1972)** *Convex Analysis.* Princeton University Press.

[14] **C.A. Floudas, P.M. Pardalos (2001)** *Encyclopedia of Optimization.* Kluwer Academic Publishers.

[15] **Apostolos Yannopoulos (2003)** *Notes on functional Analysis.* Lecture notes University of Crete (in Greek)

[16] **Wotao Yin (2016)** Coordinate Update Algorithm Short Course. UCLA Math. Lecture notes https://www.math.ucla.edu/~wotaoyin/summer2016/4_proximal.pdf

[17] Matlab code : https://web.stanford.edu/~boyd/papers/prox_algs/lasso.html#6


ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ