Introductory Remarks

“Just as nonlinear is understood in mathematics to mean not necessarily linear, we intend the term nonsmooth to refer to certain situations in which smoothness (differentiability) of the data is not necessarily postulated” [5, p.1]. Nonsmooth or, equivalently, nondifferentiable optimization tackles the problem of finding the minima (or the maxima) of real-valued functions on $\mathbb{R}^n$ in absence of differentiability hypothesis. Numerical algorithms for nonsmooth optimization aim at solving two different kind of problems: the nonsmooth convex problem and the nonsmooth nonconvex problem; therefore it is essential to know whether or not the objective function of the problem is convex.

Chapter I. Convex analysis provided not only the mathematical basis of nonsmooth convex optimization through new concepts such as the Rockafellar’s subdifferential [35], but also the basis of nonsmooth nonconvex optimization. In fact, the main tool of nonsmooth nonconvex optimization, i.e. Clarke gradient [5], is a systematic extension of the Rockafellar’s subdifferential; therefore we summarize some results and concepts of convex analysis.

Chapter II. We review the Cutting-Plane algorithm [4, 17] for minimizing convex functions and its stabilized variants, say bundle methods [16], which were introduced both by C. Lemaréchal and P. Wolfe. We report first two examples which show the inefficiency of Cutting-Plane algorithm and then we describe the penalty and the level set approaches adopted by the bundle methods.

Chapter III. Another famous method for nonsmooth convex optimization is the subgradient method [38], whose general scheme we describe. Then we present two ways to approximate the initial nonsmooth convex function by a smooth convex function;

in particular we deal with the Moreau-Yosida regularization and a new smoothing technique by Yu. Nesterov [28].

Chapter IV. We analyze some of the different definitions of differentiability, i.e. the Fréchet, the Gâuteaux and the directional differentiability. We present some concepts relative to Clarke gradient, Goldstein ϵ-subdifferential and the semismoothness. Then we describe two mean value theorems, one for directionally differentiable functions and the other for locally Lipschitzian functions, and we report a proof of Rademacher Theorem [25].

Chapter V. We present some algorithms for nonsmooth nonconvex optimization. In particular, we describe the bundle algorithms BTNC [37], NCVX [10], DC-NCVX [10] and the Gradient Sampling algorithm [3].

Chapter VI. Finally we describe the new bundle method for nonsmooth nonconvex optimization NonConvexNonSmooth (NCNS) [12]. The algorithm is based on the construction of both a lower and an upper piecewise affine approximations of the objective function. In particular, at each iteration, a search direction is calculated by solving a quadratic programming problem aiming at maximizing the difference between the lower and upper model. A proximal approach is used to guarantee convergence to a stationarity point under the hypothesis of weak semismoothness [24].

Table of Contents

Nonsmooth Convex Optimization

Chapter I

Some Elements of Convex Analysis

Introduction. The objective of the chapter is not, of course, to provide a complete description of the theory of convex functions. We give some theoretical results useful to deal with the numerical algorithms for convex programming problems.

1 Convex Functions

In this section we present some basic definitions and results related to convex analysis.

1.1 Basic Concepts

The epigraph builds a bridge between the functional language and the powerful language of the sets.

Definition 1.1.1 (Epigraph)[35] *Let* $f : A \subset \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$. *The set*

$$\operatorname{epi} f \overset{\triangle}{=} \{(x,t) \in \mathbb{R}^{n+1} : \quad x \in A,\ t \in \mathbb{R},\ t \geq f(x)\} \tag{1.1.1}$$

is called the epigraph of f.

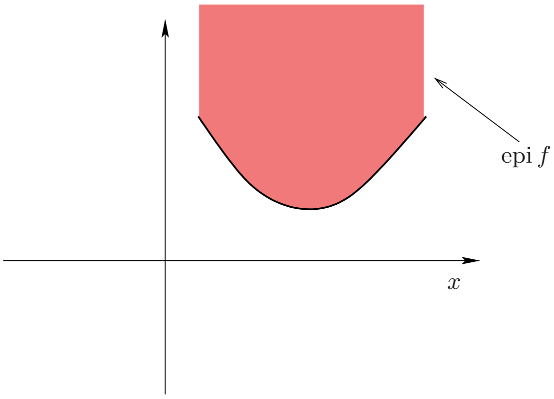


Fig. 1.1.1: Epigraph.

Definition 1.1.2 (Convex functions)[35] *Let $f : A \subset \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$. The function f is said to be convex on A, if epi f is convex as a subset of $\mathbb{R}^{n+1}$. The function f is called concave on A if $-f$ is convex on A. The function f is said to be affine on A, if it is finite, convex and concave on A.*

The following theorem provides a different definition of convex functions.

Theorem 1.1.1 [35] *Let $f : A \to \mathbb{R} \cup \{+\infty\}$, where A is a convex subset of $\mathbb{R}^n$. Then f is convex on A if and only if*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \ , \qquad \lambda \in (0,1) \tag{1.1.2}$$

for every x and y in A.

A finite convex function defined on a subset A of $\mathbb{R}^n$ can be considered as defined on the whole $\mathbb{R}^n$ by setting $f(x) = +\infty$ for $x \notin A$; for this reason we are interested in extended real-valued functions.

Definition 1.1.3 (Effective domain)[35] *Let $f : A \subset \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ be convex on A. The effective domain of f is the projection on $\mathbb{R}^n$ of the epigraph of f, i.e.*

$$\operatorname{dom} f \triangleq \{x \in \mathbb{R}^n : \quad \exists\ t,\ (x,t) \in \operatorname{epi} f\} = \{x \in A : \quad f(x) < +\infty\} \ .$$

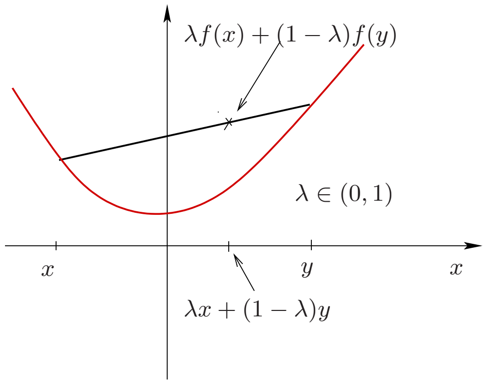


Fig. 1.1.2: A convex function

Observe that the effective domain of a convex function is a convex set.

Proposition 1.1.2 (Jensen's Inequality)[15] *Let* $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *be convex (on* $\mathbb{R}^n$*). Then*

$$f\left(\sum_{i=1}^{m} \lambda_i x_i\right) \leq \sum_{i=1}^{m} \lambda_i f(x_i) \ , \tag{1.1.3}$$

whenever $\lambda_i \geq 0, \dots, \lambda_m \geq 0$ *and* $\sum_{i=1}^{m} \lambda_i = 1$.

Proof. Take $x_i \in \operatorname{dom} f$, for $i = 1, \dots, m$. From (1.1.1), we have

$$(x_i, f(x_i)) \in \operatorname{epi} f, \quad i = 1, \dots, m \ .$$

Taking into account that $\operatorname{epi} f$ is a convex set, we obtain

$$\sum_{i=1}^{m} \lambda_i \left(x_i, f(x_i)\right) = \left(\sum_{i=1}^{m} \lambda_i x_i, \sum_{i=1}^{m} \lambda_i f(x_i)\right) \in \operatorname{epi} f \ ,$$

which implies the thesis.

□

Jensen's Inequality is "one of the most useful observations in the world" [2, p.61].

Definition 1.1.4 (Proper functions)[35] *Let* $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ *be convex. The function* f *is called proper, if its epigraph is a nonempty set and contains no vertical lines, i.e. if* $f(x) < +\infty$ *for a least one* x *and* $f(x) > -\infty$ *for all* x.

The set of proper convex functions on $\mathbb{R}^n$ taking values in the extended real axis $\mathbb{R} \cup \{+\infty\}$ is denoted by $\operatorname{Conv} \mathbb{R}^n$ [15].

1.2 Lipschitz Property of Convex Functions

Now we analyze the relationship between convex functions and Lipschitzian functions.

Proposition 1.2.1 [2] *Let $f \in \operatorname{Conv} \mathbb{R}^n$ and let $x \in \operatorname{ri} \operatorname{dom} f$. Then*

(i) *there exist $C \geq 0, r > 0$ such that*

$$|f(y)| \leq C \quad \forall\ y \in U_r(x) \overset{\triangle}{=} B_r^{(n)}(x) \cap \operatorname{aff}(\operatorname{dom} f)\ , \tag{1.2.1}$$

where $B_r^{(n)}(x)$ is the ball of radius r centered at x;

(ii) *there exist $L, \rho > 0$ such that*

$$|f(y) - f(z)| \leq L\|y - z\| \quad \forall\ y, z \in U_\rho(x)\ . \tag{1.2.2}$$

Proof. (i) For a given $x \in \operatorname{ri} \operatorname{dom} f$, there exists $r' > 0$ such that

$$U_{r'}(x) \subset \operatorname{dom} f\ .$$

Let m be the dimension of $\operatorname{aff}(\operatorname{dom} f)$ and let L be a linear subspace such that $\operatorname{aff}(\operatorname{dom} f) = L + x$. Choose m vectors, denoted as $y_1, \ldots, y_m$, forming a basis in L. Substituting $y_0 = -\sum_{i=1}^{m} y_i$ into the system

$$\left| \begin{array}{l} \sum_{i=0}^{m} \lambda_i y_i = 0 \\ \sum_{i=0}^{m} \lambda_i = 0\ , \end{array} \right. \tag{1.2.3}$$

it follows that $y_0, y_1, \ldots, y_m$ are affinely independent.

Take $\epsilon > 0$ such that $x_i \overset{\triangle}{=} \epsilon y_i + x \in U_{r'}(x)$, for $i = 0, 1, \ldots, m$. Finally construct the m-dimensional simplex with vertices $x_0, x_1, \ldots, x_m$:

$$\Delta \overset{\triangle}{=} \operatorname{co}\{x_0, \ldots, x_m\}\ .$$

Consequently

$$x = \frac{1}{m+1}\sum_{i=0}^{m} x_i,$$

which in turn implies that x is a relative interior point for Δ. It follows that there exists $r > 0$ such that

$$U_r(x) \subset \Delta \subset \operatorname{dom} f \ .$$

From Jensen's inequality (1.1.3), we have

$$f(z') \leq \sum_{i=0}^{m} \lambda_i f(x_i) \leq C' \quad \forall\ z' \in U_r(x), \tag{1.2.4}$$

where $C' \overset{\triangle}{=} \max\limits_{0\leq i\leq m} f(x_i) < \infty$.

Let z' be any point of $U_r(x)$. Then, as displayed in Fig. 1.2.1, we have

$$z'' = x - (z' - x) \in U_r(x)\ .$$

Since $x = \dfrac{z' + z''}{2}$, we have

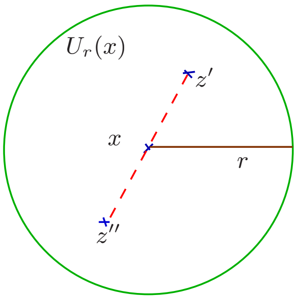


Fig. 1.2.1: $U_r(x)$.

$$f(x) = f\left(\frac{1}{2}z' + \frac{1}{2}z''\right) \leq \frac{1}{2}f(z') + \frac{1}{2}f(z''),$$

whence $f(z') \geq 2f(x) - C'$. Then, taking into account (1.2.4), (i) follows by setting $C = \max\{|C'|, |2f(x) - C'|\}$.

(ii) For any $x', x'' \in U_{\frac{r}{2}}(x)$, there exist $x^a, x^b \in \partial_{\mathrm{ri}} U_r(x)$ such that $x' \in (x^a, x'')$ and $x'' \in (x', x^b)$. Since f is finite on $U_r(x)$, from (1.1.2) we have

$$\frac{f(x') - f(x'')}{\|x' - x''\|} \leq \frac{f(x^a) - f(x'')}{\|x^a - x''\|} \tag{1.2.5a}$$

$$\frac{f(x'') - f(x')}{\|x'' - x'\|} \leq \frac{f(x^b) - f(x')}{\|x^b - x'\|} \tag{1.2.5b}$$

which yields, by using triangle inequality, the following relations

$$f(x') - f(x'') \leq \frac{f(x^a) - f(x'')}{\|x^a - x''\|} \|x' - x''\| \leq \frac{4C}{r} \|x' - x''\| \tag{1.2.6a}$$

$$f(x'') - f(x') \leq \frac{f(x^b) - f(x')}{\|x^b - x'\|} \|x'' - x'\| \leq \frac{4C}{r} \|x'' - x'\| \tag{1.2.6b}$$

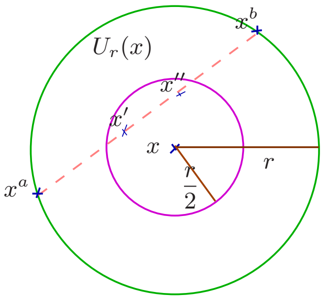


Fig. 1.2.2: $U_{\frac{r}{2}}(x)$.

By setting $\rho \triangleq \frac{r}{2}$ e $L \triangleq \frac{4C}{r}$, (ii) follows.

□

The following theorem states that a finite convex function on $\mathbb{R}^n$ is locally Lipschitzian.

Theorem 1.2.2 [2] *Let $f \in \operatorname{Conv} \mathbb{R}^n$ and let $C \subset \operatorname{ri} \operatorname{dom} f$ be a compact set. Then f is Lipschitzian on C.*

Proof. We show, by contradiction, that f is Lipschitzian on C. Suppose in fact that for each $i \in \mathbb{N}$ there exist $x_i, y_i \in C$ such that $x_i \neq y_i$ and

$$|f(x_i) - f(y_i)| > i\|x_i - y_i\|.$$

Taking into account that C is bounded, there exist two convergent subsequences

$$\{x_i\}_{i \in P}, \qquad \{y_i\}_{i \in P} \ .$$

Let

$$x = \lim_{\substack{i \to \infty \\ i \in P}} x_i, \qquad y = \lim_{\substack{i \to \infty \\ i \in P}} y_i. \tag{1.2.7}$$

Two cases can occur:

1. $x = y$. From (1.2.7), we have that there exists $\hat{\imath} \in \mathbb{N}$ such that

   $$x_i, y_i \in U_r(x), \quad \frac{|f(x_i) - f(y_i)|}{\|x_i - y_i\|} > i \qquad \forall\, i \geq \hat{\imath} \ ,$$

   which contradicts (1.2.2).

2. $x \neq y$. By virtue of Proposition 1.2.1,

   $$\lim_{\substack{i \to \infty \\ i \in P}} f(x_i) = f(x), \qquad \lim_{\substack{i \to \infty \\ i \in P}} f(y_i) = f(y) \ .$$

   Consequently we have

   $$\begin{aligned} \infty > |f(x) - f(y)| = \lim_{\substack{i \to \infty \\ i \in P}} |f(x_i) - f(y_i)| > \lim_{\substack{i \to \infty \\ i \in P}} i\|x_i - y_i\| \\ = \|x - y\| \lim_{\substack{i \to \infty \\ i \in P}} i = \infty. \end{aligned}$$

   which is a contradiction.

□

1.3 Closed Proper Convex Functions

We are usually interested in a family of convex functions with nice properties of representation. This family is formed by closed proper convex functions.

Definition 1.3.1 (Closed functions)[35] *The function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\pm\infty\}$ is said to be lower semicontinuous at x, if the condition*

$$\liminf_{y \to x} f(y) \geq f(x)$$

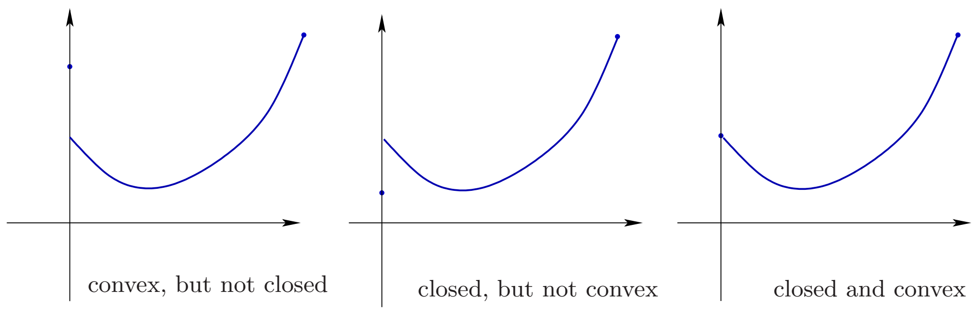


Fig. 1.3.1: Closed functions.

holds. The function f is called closed (or lower semicontinuous), if it is lower semicontinuous at every point x in $\mathbb{R}^n$ (see Fig. 1.3.1).

Remark 1.3.1 [35] *It is known that f is closed if and only if its epigraph is closed or equivalently its sublevel sets*

$$\mathcal{F}_t \overset{\triangle}{=} \{x \in \mathbb{R}^n : \quad f(x) \leq t\}, \quad \forall\, t \in \mathbb{R}$$

are closed.

The set of closed proper convex functions on $\mathbb{R}^n$ taking values in the extended real axis $\mathbb{R} \cup \{+\infty\}$ is denoted by $\overline{\text{Conv}}\mathbb{R}^n$ [15].

By considering the relationship between convex functions and convex sets, we can translate to the functional language several results related to convex sets. For example, by using Hahn-Banach Theorem (see, e.g., [15]), it is possible to prove the following basic proposition.

Proposition 1.3.1 [2] *Let $f \in \overline{\text{Conv}}\mathbb{R}^n$ and let $\mathcal{B}$ be the set of all its affine minorants, i.e.*

$$\mathcal{B} \overset{\triangle}{=} \{g : \mathbb{R}^n \to \mathbb{R} : \quad g(x) \leq f(x)\ \forall x,\ g \textit{ is affine}\}.$$

Then

$$f(x) = \sup_{g \in \mathcal{B}} g(x)$$

Moreover, if $x \in \operatorname{ri}\operatorname{dom} f$*, then*

$$f(x) = \max_{g \in \mathcal{B}} g(x) \ .$$

2 Subdifferential of Finite Convex Functions

The concept of subdifferential $\partial f(x)$ of a function f at x generalizes that of gradient $\nabla f(x)$ of f at x, in the sense that $\partial f(x)$ coincides with $\nabla f(x)$ whenever f is differentiable at x. In the definition of the subdifferential we will restrict ourselves to finite convex functions.

2.1 Support Functions

"In classic real analysis, the simplest functions are linear, in convex analysis the simplest convex functions are so-called sublinear" [15, p.195].

Definition 2.1.1 (Sublinear functions)[15] *Let* $\sigma : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ *be proper. The function* σ *is said to be sublinear, if it is convex and positively homogeneous (of degree* 1*):* $\sigma \in \operatorname{Conv} \mathbb{R}^n$ *and*

$$\sigma(tx) = t\sigma(x), \quad \forall\ x \in \mathbb{R}^n,\ t > 0\ . \tag{2.1.1}$$

It is known (see, e.g., [15]) that the function σ, participating in Definition 2.1.1, is sublinear if and only if it is subadditive, i.e.

$$\sigma(x+y) \le \sigma(x) + \sigma(y) \qquad \forall\ x, y \in \mathbb{R}^n\ ,$$

and positively homogeneous. From (2.1.1), we deduce that $\sigma(0)$ is either zero or $+\infty$. If σ is also closed, then

$$\sigma(0) \le \lim_{t \downarrow 0} \sigma(tx) = 0 \quad \forall\ x \in \operatorname{dom} f$$

and so $\sigma(0) = 0$.

Definition 2.1.2 (Support functions)[15] *Let C be a nonempty subset of $\mathbb{R}^n$. The function $\sigma_C : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ defined by*

$$\sigma_C(x) \triangleq \sup_{g \in C} g^T x \tag{2.1.2}$$

is called the support function of C.

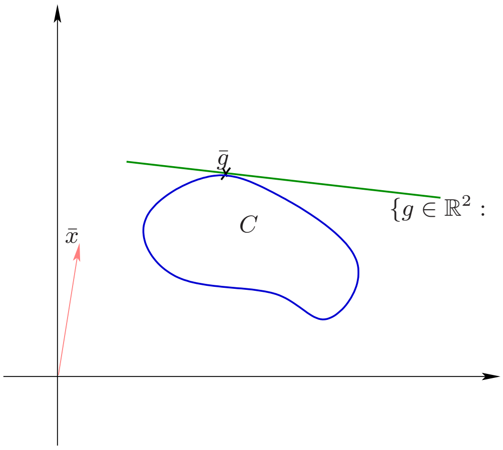

$\bar{g}$
C
$\{g \in \mathbb{R}^2 : \quad g^T\bar{x} = \sigma_C(\bar{x})\}$
$\bar{x}$

Fig. 2.1.1: The geometrical meaning of the support function.

For a given nonempty set C, the supremum in (2.1.2) may be finite or not finite, achieved on C or not (see Fig. 2.1.1).

Proposition 2.1.1 [15] *The support function $\sigma_C : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ of a set C is finite everywhere if and only if C is bounded.*

Proposition 2.1.2 [15] *The support function $\sigma_C : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ of a set C is closed and sublinear.*

Proof. Let $x \mapsto l_g(x) = g^T x : \mathbb{R}^n \to \mathbb{R}$. Since $\operatorname{epi} \sigma_C = \bigcap_{g \in C} \operatorname{epi} l_g$, it follows that σ_C is convex and closed. Moreover, $\sigma_C(0) = 0$ implies that σ_C is proper. It is clear that f is positively homogeneous and therefore we obtain the result.

□

The closed convex hull of any nonempty set can be express through its support function. In fact we have the following result.

Theorem 2.1.3 [15] *Let $C \subset \mathbb{R}^n$ be nonempty and let σ_C be the support function of C. Then*

$$\overline{\text{co}}\ C = \left\{ g \in \mathbb{R}^n : \quad g^T x \leq \sigma_C(x), \quad \forall\ x \in \mathbb{R}^n \right\} \tag{2.1.3}$$

Proof.

Considering that the closure of the convex hull of a set coincides with its closed convex hull (see, e.g., [15]), the property $\sigma_C = \sigma_{\overline{\text{co}} C}$ is a consequence of the continuity and convexity of the function $x \mapsto g^T x$. Thus, from Definition 2.1.2, we have

$$\overline{\text{co}}\ C \subset \left\{ g \in \mathbb{R}^n : \quad g^T x \leq \sigma_C(x), \quad \forall\ x \in \mathbb{R}^n \right\}$$

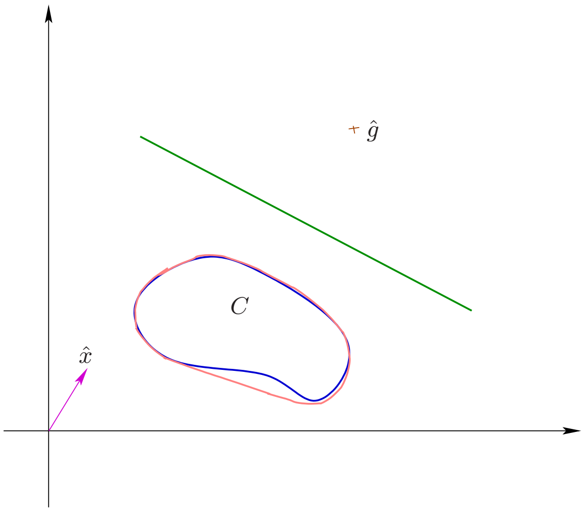


Fig. 2.1.2: An application of Hahn-Banach Theorem.

If $\hat{g} \notin \overline{\text{co}}\ C$, by Hahn-Banach Theorem (see, e.g., [15]), there exists $\hat{x} \in \mathbb{R}^n$, as displayed in Fig. 2.1.2, such that

$$\hat{g}^T \hat{x} > \sigma_{\overline{\text{co}} C}(\hat{x})$$

Consequently we have

$$\overline{\text{co}}\ C \supset \left\{ g \in \mathbb{R}^n : \quad g^T x \leq \sigma_C(x), \ \ \forall\ x \in \mathbb{R}^n \right\}$$

□

Theorem 2.1.4 [15] *Let $\sigma : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be closed and sublinear. Then σ is the support function of the nonempty closed convex set*

$$C = \left\{ g \in \mathbb{R}^n : \quad g^T x \leq \sigma(x), \ \ \forall\ x \in \mathbb{R}^n \right\}$$

2.2 Subdifferential of Finite Convex Functions

Now we prove that the directional derivative is a particular sublinear function.

Proposition 2.2.1 [15] *Let $f \in \operatorname{Conv} \mathbb{R}^n$ be finite. For a fixed $x \in \mathbb{R}^n$, the function*[1] *$d \to f'(x, d)$ on $\mathbb{R}^n$ is finite sublinear.*

Proof. Let $d_1, d_2 \in \mathbb{R}^n$ and $\lambda \in (0, 1)$. Since f is convex, we have

$$\begin{aligned} t \in \mathbb{R}, \quad f(x + t(\lambda d_1 {+} (1 - \lambda) d_2)) - f(x) &= f\left(\lambda(x + td_1) + (1 - \lambda)(x + td_2)\right) - f(x) \\ &\leq \lambda[f(x + td_1) - f(x)] + (1 - \lambda)[f(x, td_2) - f(x)] \ . \end{aligned}$$

Dividing by $t > 0$ and by passing to limit as $t \downarrow 0$, we have

$$f'(x, \lambda d_1 + (1 - \lambda) d_2) \leq \lambda f'(x, d_1) + (1 - \lambda) f'(x, d_2) \ . \tag{2.2.1}$$

Consequently $d \to f'(x, d)$ is convex. This function is also positively homogeneous; in fact for every $\lambda > 0$ we have

$$\begin{aligned} f'(x, \lambda d) &= \lim_{t \downarrow 0} \frac{f(x + t\lambda d) - f(x)}{t} \\ &= \lim_{\tau \downarrow 0} \lambda \frac{f(x + \tau d) - f(x)}{\tau} \\ &= \lambda f'(x, d) \ . \end{aligned} \tag{2.2.2}$$

[1] $f'(x, d)$ is the directional derivative of f at x in the direction d.

Finally, from Theorem 1.2.2, we have

$$|f'(x,d)| = \lim_{t\downarrow 0}\left|\frac{f(x+td)-f(x)}{t}\right| \le L\|d\| \qquad \forall\ d \in \mathbb{R}^n, \tag{2.2.3}$$

where L is the Lipschitz constant of f around x and the thesis follows.

□

After we have proved that $d \to f'(x,d)$ is finite and sublinear, we can immediately apply Theorem 2.1.4 and extrapolate the following definition.

Definition 2.2.1 (Subdifferential)[15] *Let $f \in \operatorname{Conv}\mathbb{R}^n$ be finite and let x any point of $\mathbb{R}^n$. The subdifferential of f at x, denotes as $\partial f(x)$, is the nonempty compact convex subset of $\mathbb{R}^n$ whose support function is $d \to f'(x,d)$, i.e.*

$$\partial f(x) \overset{\triangle}{=} \left\{ g \in \mathbb{R}^n : \quad g^T d \le f'(x,d), \ \ \forall\ d \in \mathbb{R}^n \right\} \tag{2.2.4}$$

From another point of view the subdifferential is defined as follows.

Definition 2.2.2 [15] *Let $f \in \operatorname{Conv}\mathbb{R}^n$ be finite and let x any point of $\mathbb{R}^n$. The subdifferential of f at x is the set of vectors $g \in \mathbb{R}^n$ satisfying*[2]

$$f(y) \ge f(x) + g^T(y-x) \ \ \forall\ y \in \mathbb{R}^n \ . \tag{2.2.5}$$

A vector $g \in \partial f(x)$ is said to be a subgradient of f at x.

The two definitions of the subdifferential are equivalent (see [35]) and we remark that the subdifferential is a generalization of the concept of Fréchet differentiability. To see this it is sufficient to prove that if f is differentiable at x, then the subdifferential is the singleton

$$\partial f(x) = \{\nabla f(x)\} \ .$$

In fact given a vector $g \in \partial f(x)$, we have

$$f(x+td) - f(x) \ge t g^T d \ \ \forall\ d \in \mathbb{R}^n \ ,$$

whence $\nabla f(x)^T d \ge g^T d$ for all $d \in \mathbb{R}^n$. Last inequality is possible if and only if $g = \nabla f(x)$.

[2] (2.2.5) is called the “subgradient inequality” [35].

Proposition 2.2.2 [15] *Let $f \in \operatorname{Conv}\mathbb{R}^n$ be finite and let x any point of $\mathbb{R}^n$. Then the subdifferential mapping of f is outer semicontinuous, i.e.*

$$\forall\ \epsilon > 0, \quad \exists\ \delta > 0: \quad y \in B_\delta^{(n)}(x) \Rightarrow \partial f(y) \subset \partial f(x) + B_\epsilon^{(n)}(0). \tag{2.2.6}$$

Moreover, for a given $d \in \mathbb{R}^n$ the function $x \mapsto f'(x,d)$ is upper semicontinuous, i.e.

$$f'(x,d) = \limsup_{y \to x} f'(y,d) \quad \forall\ x \in \mathbb{R}^n \tag{2.2.7}$$

Theorem 2.2.3 [15] *Let $f \in \operatorname{Conv}\mathbb{R}^n$ be finite and let x any point of $\mathbb{R}^n$. Then, for every $x \in \mathbb{R}^n$ we have*

$$\partial f(x) = \text{co}\ \{\lim_{i\to\infty} \nabla f(x_i): \quad \lim_{i\to\infty} x_i = x, \quad x_i \notin G_f\},$$

where $G_f \overset{\triangle}{=} \{x \in \mathbb{R}: \quad f$ is not differentiable at $x\}$

Proof. Consider the set

$$\gamma f(x) \overset{\triangle}{=} \{\lim_{i\to\infty} \nabla f(x_i): \quad \lim_{i\to\infty} x_i = x, \quad x_i \notin G_f\}$$

From (2.2.3), it follows that ∇f is bounded around x. Thus $\gamma f(x)$ is a nonempty bounded set. Let $\{s_k\}_{k\in\mathbb{N}}$ be a sequence such that $s_k = \lim_{x_{ki}\to x} \nabla f(x_{ki})$. Thus we have

$$\lim_{k\to\infty} s_k = \lim_{k\to\infty} \lim_{i\to\infty} \nabla f(x_{ki}) \in \gamma f(x).$$

which implies that $\gamma f(x)$ is a closed set. Taking into account that the convex hull of a compact set is a compact set (see, e.g., [15]), we have that $\text{co}\,\gamma f(x)$ is a compact set.

Moreover (see [15]), by Proposition 2.2.2, it follows that

$$\gamma f(x) \subset \text{co}\,\gamma f(x) \subset \partial f(x)\ .$$

Let $\sigma_{\gamma f(x)}$ the support function of $\gamma f(x)$. By Theorem 2.1.3, we have

$$\text{co}\,\gamma f(x) = \{g: \quad g^T d \leq \sigma_{\gamma f(x)}(d)\ \forall d\ \in \mathbb{R}^n\}\ .$$

To show that $\text{co}\,\gamma f(x) \supset \partial f(x)$, we prove by contradiction that

$$f'(x,d) \leq \sigma_{\gamma f(x)}(d) \quad \forall\ d \in \mathbb{R}^n$$

Suppose in fact that there exist $d^* \in \mathbb{R}^n$ and $\bar{\epsilon} > 0$ such that

$$\sigma_{\gamma f(x)}(d^*) < f'(x, d^*) - \bar{\epsilon} \; . \tag{2.2.8}$$

From (2.1.2), for all $d \in \mathbb{R}^n$ we have

$$\sigma_{\gamma f(x)}(d) = \limsup_{\substack{y \to x \\ y \in G_f}} \nabla f(y)^T d \; .$$

Or equivalently, for every $d \in \mathbb{R}^n$, $\epsilon > 0$ there exists $\delta > 0$ such that

$$y \in G_f, \; \|y - x\| \leq \delta \Rightarrow \nabla f(y)^T d \leq \sigma_{\gamma f(x)}(d) + \frac{\epsilon}{2} \; .$$

Consequently, from (2.2.8), we have that there exists $\bar{\delta} > 0$ such that

$$y \in G_f, \; \|y - x\| \leq \bar{\delta} \Rightarrow \nabla f(y)^T d^* < f'(x, d^*) - \frac{\bar{\epsilon}}{2}.$$

Let $B_{\bar{\delta}}(x, d^*) \overset{\triangle}{=} \{y \in B_{\bar{\delta}}^{(n)}(x) : \; y^T d^* = 0\}$ and $L_y \overset{\triangle}{=} \{y + td^* : \; t \in \mathbb{R}\}$. By virtue of both Fubini's Theorem (see [34]) and Rademacher Theorem IV.4.2.2, we derive that the linear measure of $L_y \cap G_f$ is zero for almost all $y \in B_{\bar{\delta}}(x, d^*)$.

Let y^* be any point of $B_{\bar{\delta}}(x, d^*)$ such that $\lambda_1(L_{y^*} \cap G_f) = 0$.
Consider $\phi(t) \overset{\triangle}{=} f(y^* + td^*)$. Since ϕ is absolutely continuous (see [34]), we have

$$y^* + td^* \in B_{\bar{\delta}}(x, d^*), \qquad \frac{f(y^* + td^*) - f(x)}{t} = \frac{\phi(t) - \phi(0)}{t} < \frac{\int_0^t \left(f'(x, d^*) - \frac{\bar{\epsilon}}{2}\right) \mathrm{d}\tau}{t}$$
$$= f'(x, d^*) - \frac{\bar{\epsilon}}{2},$$

By continuity of f, we have

$$f'(y, d^*) \leq f'(x, d^*) - \frac{\bar{\epsilon}}{2} \quad \forall \; y \in B_{\bar{\delta}}^{(n)}(x, d^*)$$

which contradicts (2.2.7).

□

Chapter II

Bundle Methods

Introduction. In this chapter we deal with the bundle methods. These efficient algorithms were introduced both by C. Lemaréchal and P. Wolfe in the Seventies [16]. The bundle methods are stabilized versions of the Cutting-Plane algorithm. Consequently, before discussing the bundle methods, we present the Cutting-Plane algorithm, introduced by both Cheney and Goldestein [4] (1959) and Kelley [17] (1960).

1 Cutting-Plane Algorithm

We consider the following problem

$$(P) \begin{cases} \min f(x) \\ x \in \mathcal{C}, \end{cases} \tag{1.0.1}$$

where $f \in \operatorname{Conv} \mathbb{R}^n$ is finite and $\mathcal{C}$ is a compact convex set. The set $\mathcal{C}$ is introduced to overcome problems of convergence and the problem (P) is not the standard constrained problem. We assume the existence of an oracle which, given any point $x \in \mathcal{C}$, computes both the objective function value $f(x)$ and a subgradient $g \in \partial f(x)$. We assume also that a starting point belonging to $\mathcal{C}$, say x_1, is available.

1.1 The Algorithm

The main idea of the Cutting-Plane consists in minimizing a piecewise affine approximation of the objective function and such lower approximation is more and more enriched during the execution of the algorithm. Let $x_1, \ldots, x_k$ be k points and $g_j \in \partial f(x_j)$ any subgradient of f at x_j, with $j \in 1, \ldots, k$. Then we construct the following polyhedral approximation of f

$$\check{f}_k(x) \overset{\triangle}{=} \max_{1 \leq j \leq k} \left\{ f(x_j) + g_j^T(x - x_j) \right\}.$$

This function $\check{f}_k$ is called the "cutting-plane function" associated with points $x_1, \ldots, x_k$.

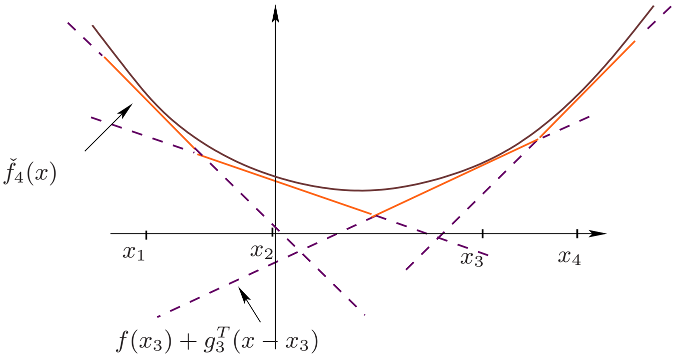


Fig. 1.1.1: Cutting-plane function.

We remark (Proposition I.1.3.1) that f is the maximum of all its linearizations, namely

$$f(x) = \max_{x_j \in \mathbb{R}^n} \left\{ f(x_j) + g_j^T(x - x_j) \right\}.$$

Consequently $\check{f}_k$, for $k \geq 1$, is an under-estimate of f.

The following lemma establishes some results on cutting-plane functions.

Lemma 1.1.1 *For all $x \in \mathbb{R}^n$, we have*

(i) $\check{f}_j(x) \leq \check{f}_k(x) \quad \forall\ 1 \leq j \leq k$;

(ii) $\check{f}_k(x) \leq f(x) \quad \forall\ k \geq 1$;

(iii) $\check{f}_k(x_j) = f(x_j) \quad \forall\ 1 \leq j \leq k.$

Proof. The assertion (i) follows directly from definition of the cutting-plane function. Moreover the subgradient inequality implies (ii). In fact we have

$$f(x) \geq f(x_j) + g_j^T(x - x_j) \quad \forall\ 1 \leq j \leq k.$$

The property (iii) is a consequence of the convexity of f.

□

Finally we can present the Cutting-Plane algorithm.

Algorithm 1.1.2 (Cutting-Plane algorithm) [16]

Step 0. *Choose* $\epsilon \geq 0$ *and* $x_1 \in \mathcal{C}$. *Set* $k := 1$.

Step 1. *Solve the* k^{th} *"cutting-plane" problem*

$$(P_k) \begin{cases} \min \ \check{f}_k(x) \\ \qquad x \in \mathcal{C} \end{cases}$$

to obtain a solution $\bar{x}$.

Step 2. *Set* $x_{k+1} := \bar{x}$, *evaluate* $f(x_{k+1})$ *and calculate a subgradient* $g_j \in \partial f(x_j)$.

Step 3. *If*

$$f(x_{k+1}) \leq \check{f}_k(x_{k+1}) + \epsilon,$$

then stop. Otherwise set $k := k+1$ *and return to step* 1.

Of course the subproblem (P_k) is equivalent to following simpler problem

$$\begin{cases} \min \ v \\ \quad v \geq f(x_j) + g_j^T(x - x_j) \quad \forall\ j = 1, \ldots k \\ \quad x \in \ \mathcal{C} \end{cases}$$

We remark that, since $\mathcal{C}$ is bounded, the cutting-plane function is bounded from below. To solve (P_k), we must maintain the information about the points previously generated; so, at k^{th} iteration of the algorithm, the "bundle" B_k denotes the set of triplets

$$\{(x_j, f(x_j), g_j) : \quad 1 \leq j \leq k\}.$$

1.2 Convergence

Now we prove the convergence of Algorithm 1.1.2.

Theorem 1.2.1 *Let f_c^* the optimum value of (P). Then we have*

(i) *If $\epsilon > 0$, then the algorithm stops in a finite number of iterations at a point x_{k+1} satisfying the condition*

$$f(x_{k+1}) \leq f_c^* + \epsilon;$$

(ii) *If $\epsilon = 0$, then both $\min\limits_{0 \leq i \leq k} f(x_{i+1})$ and $\check{f}_k(x_{k+1})$ tend to f_c^* when $k \to \infty$.*

Proof. (i) If stopping condition at step 3 is satisfied, then we have

$$f(x_{k+1}) \leq \check{f}_k(x_{k+1}) + \epsilon \leq f_c^* + \epsilon.$$

Suppose that the algorithm loops infinitely many times. Then for all indices k we have

$$f(x_{k+1}) - \epsilon > \check{f}_k(x_{k+1}) \geq f(x_j) + g_j^T(x_{k+1} - x_j) \quad \forall\ 1 \leq j \leq k.$$

Hence, by Lipschitz property (Theorem I.1.2.2) and letting L be the Lipschitz constant of f in $\mathcal{C}$, we have

$$-\epsilon \geq -2L\|x_{k+1} - x_j\| \quad \forall\ 1 \leq j \leq k,$$

which is a contradiction, since C is compact.

(ii) For all $\delta > 0$, there exists an index $\bar{k}$ such that at step 2

$$f(x_{\bar{k}+1}) \leq \check{f}_{\bar{k}}(x_{\bar{k}+1}) + \delta$$

is satisfied (see the above proof of (i)). That is, for all $\delta > 0$, there exists an index $\bar{k}$ such that

$$0 \leq \min_{0 \leq i \leq k} f(x_{i+1}) - \check{f}_k(x_{k+1}) \leq \delta \quad \forall\ k \geq \bar{k}.$$

Taking into account that $\{\min\limits_{0 \leq i \leq k} f(x_{i+1})\}$ and $\{\check{f}_k(x_{k+1})\}$ are monotone sequences bounded by f_c^*, respectively, from below and above, i.e.

$$\check{f}_k(x_{k+1}) \leq f_c^* \leq \min_{0 \leq i \leq k} f(x_{i+1}) \ ,$$

the thesis follows.

□

1.3 Instability of Cutting-Plane

The Cutting-Plane algorithm is unstable and its numerical performance is intolerably low. We discuss these facts by means of two examples.

Example 1.3.1 [16] *Consider the function*

$$f(x) = \frac{1}{2}x^2 \ .$$

The minimum is at $x^* = 0$. *To prove the instability of the Cutting-Plane algorithm, we consider two iterates* $x_1 = 1$ *and* $x_2 = -\epsilon$. *The cutting-plane approximation of* f*, associated with these points, is :*

$$\check{f}_2(x) = \max\left\{x - \frac{1}{2}, -\epsilon x - \frac{1}{2}\epsilon^2\right\}.$$

For a small $\epsilon > 0$, x_2 *is close to point* x^**, while the new iterate* $x_3 = \frac{1}{2} - \frac{1}{2}\epsilon$ *is far from* x^**.*

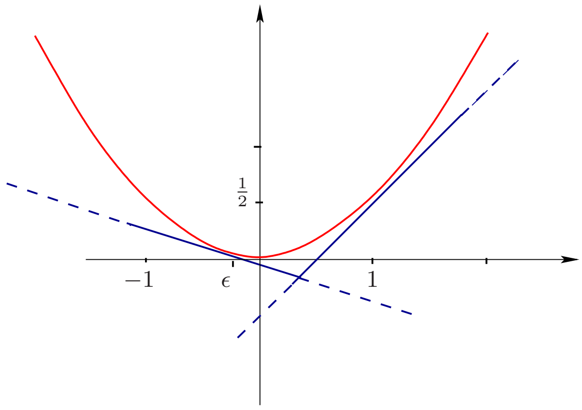


Fig. 1.3.1: Instability of Cutting-Plane.

The next example, proposed by A. Nemirovskij, proves that the Cutting-Plane algorithm could require a large number of iterations.

Example 1.3.2 [16] *Let $\epsilon \in (0, \frac{1}{2})$. Consider the following minimization problem*

$$(\mathcal{P}) \begin{cases} \min f(y,\eta) \\ (y,\eta) \in \mathcal{C} \ , \end{cases}$$

where the objective function is

$$(y,\eta) \mapsto f(y,\eta) = \max\{|\eta|, -1 + 2\epsilon + \|y\|\} : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$$

and $\mathcal{C}$ is the unit ball $B^{(n+1)}$.

The function f attains its minimum on $\mathcal{C}$ at all points of the set

$$\operatorname*{Argmin}_{\mathcal{C}} f = \left\{(y,0) : \quad y \in B^{(n)}_{1-2\epsilon}\right\} \ ,$$

with minimum value $f^*_c = 0$.

Let $x_1 \overset{\triangle}{=} (y_1, \eta_1) = (0,1)$ be the starting point. Therefore we have $f(x_1) = 1$ and $\check{f}_1(x) = \eta$. Then we solve the 1^{st} cutting-plane problem

$$(P_1) \begin{cases} \min \ v \\ \quad v \geq \eta \\ \quad \|y\|^2 + \eta^2 \leq 1 \ . \end{cases}$$

Its minimal value is $v^* = -1$, obtained at the point

$$(v^*, y^*, \eta^*) = (-1, 0, -1) \ .$$

Consequently, $x_2 = (0,-1)$. Thus we have $f(x_2) = 1$ and $\check{f}_2(x) = |\eta|$. Then we solve the 2^{nd} cutting-plane problem

$$(P_2) \begin{cases} \min \ v \\ \quad v \geq |\eta| \\ \quad \|y\|^2 + \eta^2 \leq 1 \end{cases}$$

Its minimal value is $v^* = 0$, obtained at all points of the set

$$\{(v^*, y^*, \eta^*) : \quad v^* = \eta^* = 0, \ \|y^*\| \leq 1\}$$

Let $x_3 \in \{(y_3, 0) : \quad \|y_3\| = 1\}$*. Therefore we have* $f(x_3) = 2\epsilon$ *and* $\check{f}_3(x) = \max\{|\eta|, 2\epsilon + y_3^T(y - y_3)\}$*. Then we solve the* 3^{rd} *cutting-plane problem*

$$(P_3) \begin{cases} \min \ v \\ \quad v \geq |\eta| \\ \quad v \geq 2\epsilon + y_3^T(y - y_3) \\ \quad \|y\|^2 + \eta^2 \leq 1 \end{cases}$$

Its minimal value is $v^* = 0$*, obtained at all points of the set*

$$\Big\{(v^*, y^*, \eta^*) : \quad v^* = \eta^* = 0, \ \|y^*\| \leq 1, \ {y^*}^T y_3 \leq 1 - 2\epsilon\Big\}$$

Let $x_4 \in \{(y_4, 0) : \quad \|y_4\| = 1, \ y_4^T y_3 \leq 1 - 2\epsilon\}$*. Therefore we have* $f(x_4) = 2\epsilon$*. We remark that* $v^* = \eta^* = 0$*, for* $k \geq 2$*. In fact the following formula holds:*

$$0 = \check{f}_2(x_3) \leq \check{f}_k(x_{k+1}) \leq f_c^* = 0 \qquad k \geq 2.$$

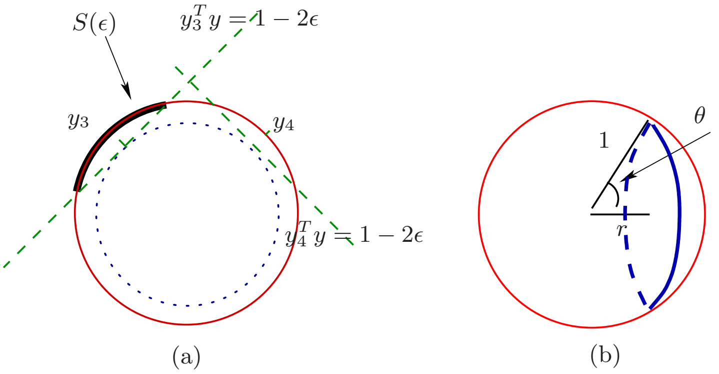


Fig. 1.3.2: Cuts and surfaces in the unit ball.

Before all the vectors of norm one are eliminated by successive cuts (see Fig. 1.3.2(a)), we can take

$$x_{i+1} \in \Big\{(y_{i+1}, 0) : \quad \|y_{i+1}\| = 1, \quad y_{i+1}^T y_p \leq 1 - 2\epsilon, \ p = 3, \ldots, i\Big\}$$

where x_{i+1} *is a minimizer of* (P_i)*.*

It is known that the area of the sphere $S_r^{(n)}(0)$ is $r^{n-1}S_n$, where S_n is the area of the unit sphere $S^{(n)}$. Furthermore the area of the infinitesimal ring in $\mathbb{R}^n$ at distance r of the origin, displayed in Fig. 1.3.2(b), is

$$S_{n-1}(\sqrt{1-r^2})^{n-2}\mathrm{dr} = \mathrm{S_{n-1}(\sin\theta)^{n-1}d}\theta.$$

We define $S(\epsilon)$ as the area of i^{th} "cap", shown in Fig. 1.3.2(a), i.e.

$$\left\{y \in \mathbb{R}^n : \quad \|y\| = 1, \quad y^T y_i \geq 1 - 2\epsilon\right\}.$$

At least $2 + \dfrac{S_n}{S(\epsilon)}$ iterations will occur, before the stopping condition at step 3 of Algorithm 1.1.2 can be satisfied. Setting $\theta_\epsilon \overset{\triangle}{=} \cos^{-1}(1-2\epsilon)$, we have

$$\begin{aligned} S(\epsilon) =& S_{n-1}\int_0^{\theta_\epsilon}(\sin\theta)^{n-1}\mathrm{d}\theta \leq S_{n-1}\int_0^{\theta_\epsilon}(\theta)^{n-1}\mathrm{d}\theta = S_{n-1}\frac{1}{n}(\theta_\epsilon)^n \\ S_n =& 2S_{n-1}\int_0^{\frac{\pi}{2}}(\sin\theta)^{n-1}\mathrm{d}\theta \geq 2S_{n-1}\int_0^{\frac{\pi}{2}}(\sin\theta)^{n-1}\cos\theta\mathrm{d} = \frac{2}{n}S_{n-1}. \end{aligned}$$

Tacking into account that $\theta_\epsilon \simeq 2\sqrt{\epsilon}$ for a given small ϵ, we have

$$\frac{S_n}{S(\epsilon)} \simeq \frac{2}{(2\sqrt{\epsilon})^n} \; .$$

2 Stabilized Variants of Cutting-Plane

Consider the minimization problem (P) mentioned in §1 with $\mathcal{C} = \mathbb{R}^n$. Some refinements of the Cutting-Plane algorithm have been studied since 1975. These algorithms are so-called bundle methods. We present three variants of bundle methods depending on how we calculate the next iterate:

- Cutting-plane with stabilization by penalty. This is also the most popular bundle method. To stabilize the algorithm, the "stability center" y_k has been introduced; in particular y_k is the current estimate of the minimum of f. The next iterate is calculated as

$$x_{k+1} := \operatorname*{argmin}_{x\in\mathbb{R}^n}\left\{\check{f}_k(x) + \frac{1}{2}\rho\|x - y_k\|^2\right\}, \qquad (2.0.1)$$

where ρ, the weight of quadratic term, is the current "proximity parameter" (or "penalty"). If f evaluated at x_{k+1} turns out to be "sufficiently decreased" with respect to its value at the stability center y_k, then we update the stability center, i.e. $y_{k+1} = x_{k+1}$ (descent step). Otherwise the stability center is unchanged, i.e $y_{k+1} = y_k$ (null step).

- Cutting-plane with "trust region". The next iterate is calculated as

$$x_{k+1} := \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \check{f}_k(x), \|x - y_k\| \leq \tau_k \right\} \ ,$$

that is $\check{f}_k(x)$ is minimized on a ball of center y_k. The stability center is updated using the same logic as the previous case.

- Cutting-plane with level-stabilization. In this approach the next iterate is computed as

$$x_{k+1} := \operatorname*{argmin}_{x \in \mathbb{R}^n} \left\{ \frac{1}{2} \|x - y_k\|^2, \check{f}_k(x) \leq l_k \right\}, \tag{2.0.2}$$

where l_k is the current level; in particular l_k is the "current estimate" of $\min\limits_{\mathbb{R}^n} f$. In this case the stability center is simply the last iterate.

In [9] it is proved that the above approaches are substantially equivalent. Now we restrict our attention to the cutting-plane with stabilization by penalty and the cutting-plane with level-stabilization.

2.1 Cutting-Plane with Stabilization by Penalty

Now we consider the penalization point of view. The problem (2.0.1) is equivalent to the following quadratic problem

$$\left\{ \begin{array}{l} \min \ r + \frac{1}{2}\rho \|x - y_k\|^2 \\ \quad r \geq f(x_j) + g_j^T (x - x_j) \quad \forall \ j = 1, \ldots k \ . \end{array} \right. \tag{2.1.1}$$

To implement an efficient method, for all index j we introduce the quantity

$$\alpha_j^k \triangleq f(y_k) - [f(x_j) + g_j^T (y_k - x_j)],$$

that is the j^{th} linearization error, i.e. it is the difference between the actual value of f at y_k and the linear expansion of f generated at x_j and evaluated at y_k.

Set $v \stackrel{\triangle}{=} r - f(y_k)$ and $d \stackrel{\triangle}{=} x - y_k$, so that the problem (2.1.1) becomes

$$(QP_\rho)\begin{cases} \min\limits_{v,d} \ v + \dfrac{1}{2}\rho\|d\|^2 \\ \quad v \geq g_j^T d - \alpha_j^k \quad \forall\ j \in I_k, \end{cases} \tag{2.1.2}$$

where $I_k \stackrel{\triangle}{=} \{1, \dots, k\}$ are the points previously generated. By duality this is equivalent to finding multiplier vector $\lambda(\rho)$ that solve the quadratic problem

$$(DP_\rho)\begin{cases} \min\limits_{\lambda \geq 0} \ \dfrac{1}{2\rho}\,\|G\lambda\|^2 + \lambda^T \alpha^k \\ \quad e^T\lambda = 1\ , \end{cases}$$

where G is the matrix whose columns are the vectors g_j, $j \in I_k$. Analogously, the terms α_j^k, $j \in I_k$, are grouped into the vector of appropriate dimension α^k.

We indicate by $(d(\rho), v(\rho))$ and $\lambda(\rho)$, respectively, the optimal solutions of (QP_ρ) and (DP_ρ) so that the role of ρ is emphasized. The following primal-dual relations hold:

$$d(\rho) = -\frac{1}{\rho} G\lambda(\rho) \tag{2.1.3a}$$

$$v(\rho) = -\frac{1}{\rho}\,\|G\lambda(\rho)\|^2 - \lambda(\rho)^T\alpha^k \tag{2.1.3b}$$

Theorem 2.1.1 *Let $\lambda(\rho)$ be the optimal solution of (QP_ρ) with*

$$\|G\lambda(\rho)\| \leq \epsilon \quad \text{and} \quad \lambda(\rho)^T\alpha^k \leq \epsilon.$$

Then y_k is ϵ-optimal, i.e.

$$f(y_k) \leq f(x) + \epsilon\|x - y_k\| + \epsilon \quad \forall\ x \in \mathbb{R}^n.$$

Proof. Choose any $x \in \mathbb{R}^n$. Then, since the subgradient inequality holds, we have:

$$f(x) \geq f(y_k) + g_j^T(x - y_k) - \alpha_j^k \quad \forall\ j \in I_k.$$

Consequently we have

$$f(y_k) \leq f(x) - \left(\sum_{j=1}^{k} \lambda_j(\rho) g_j\right)^T (x - y_k) + \sum_{j=1}^{k} \lambda_j(\rho)\alpha_j^k \leq f(x) + \epsilon\|x - y_k\| + \epsilon.$$

□

The storage capacity is finite for any computer, thus it is not possible to augment the bundle size indefinitely. To cope with this difficulty, we define the aggregate subgradient $g_p^k \triangleq \sum_{j=1}^k \lambda_j(\rho) g_j$ and the aggregate linearization error $\alpha_p^k \triangleq \sum_{j=1}^k \lambda_j(\rho)\alpha_j^k$. The following problems shave the same optimal solution

$$(QP_\rho)\begin{cases} \min\ v + \frac{1}{2}\rho\|d\|^2 \\ v \geq g_j^T d - \alpha_j^k \quad \forall\ j \in I_k \end{cases} \qquad (QP_\rho^a)\begin{cases} \min\ v + \frac{1}{2}\rho\|d\|^2 \\ v \geq {g_p^k}^T d - \alpha_p^k \\ v \geq g_j^T d - \alpha_j^k \quad \forall\ j \in \bar{I}_k, \end{cases}$$

where $\bar{I}_k$ is an arbitrary subset possibly empty of I_k. Let $\bar{l}$ be the upper threshold on the bundle capacity. If $|I_k| \geq \bar{l}$, then we delete a part of the bundle, i.e. $I_{k+1} = \bar{I}_k \cup p$, where p is the index of the aggregate bundle element p.

Hence we can present a typical bundle method with stabilization by penalty.

Algorithm 2.1.2 [8]

Step 0. *Select the starting point x_1. Choose the stopping parameter ϵ, the maximal bundle size $\bar{l}$ and the descent parameter $m \in (0,1)$. Put $y_1 := x_1$. Set the initial bundle*[1] *$B_1 \triangleq \{(0, g_1)\}$ and $k := 1$.*

Step 1 *(main computation and stopping test). Select a value ρ of the proximity parameter and solve either (QP_ρ) or (DP_ρ). If $\rho\|d_k\| \leq \epsilon$ and $-v_k - \rho\|d_k\|^2 \leq \epsilon$, then stop: y_k is ϵ-optimal. Else put $x_{k+1} := y_k + d_k$ and calculate g_p^k and α_p^k.*

Step 2 *(descent test). Evaluate $f(x_{k+1})$ and calculate $g_{k+1} \in \partial f(x_{k+1})$; if the descent test*

$$f(x_{k+1}) \leq f(x_k) + mv^k$$

is not satisfied put $y_{k+1} := y_k$ and go to step 4.

Step 3 (descent or serious step). Change the stability center: $y_{k+1} := x_{k+1}$ and update the linearization error for all index j of the bundle B_k:

$$\alpha_j^{k+1} := \alpha_j^k + f(y_{k+1}) - f(y_k) + g_j^T(y_k - y_{k+1})$$

[1] The set of triplets B_k is replaced by the set of pairs $\{(\alpha_j^k, g_j) : \quad j \in I_k\}$

Step 4 (managing the bundle size) If $|I_k| = \bar{l}$, then delete at least 2 elements from the bundle and add the element (α_p^k, g_p^k) to the bundle.
Step 5 Set $l = |I_k|$ and

$$\alpha_{l+1}^{k+1} := \begin{cases} 0 & \text{in case of serious step} \\ f(y_{k+1}) - [f(x_{k+1}) + g_{k+1}^T(y_{k+1} - x_{k+1})] & \text{in case of null step} \end{cases}$$

Update the bundle $B_{k+1} := B_k \cup \{\alpha_{l+1}^{k+1}, g_{k+1}\}$, replace k by $k+1$ and return to step 1.

See [16] for convergence properties of the bundle methods.

2.2 Cutting-Plane with Level-Stabilization

We restrict ourselves to the problem of minimizing a finite convex function f on a nonempty compact convex set[2] $\mathcal{C}$. The core of the algorithm is the solution of the problem

$$x_{k+1} := \operatorname*{argmin}_{x \in \mathcal{C}} \left\{ \frac{1}{2}\|x - x_k\|^2, \check{f}_k(x) \leq l_k \right\}.$$

L denotes the Lipschitz constant of f in $\mathcal{C}$, D denotes the diameter of $\mathcal{C}$ with respect to the Euclidian norm, $\pi(x, \mathcal{C})$ denotes the unique point of $\mathcal{C}$ closest to x, finally f_c^* denotes the minimum value of f over $\mathcal{C}$.

Moreover we define the quantity

$$\epsilon(x) \triangleq \begin{cases} +\infty & x \notin \mathcal{C} \\ f(x) - \min\limits_{\mathcal{C}} f & x \in \mathcal{C}, \end{cases}$$

Algorithm 2.2.1 [19]
Step 0 *(initialization). Choose* $\lambda \in (0,1)$. *Select a starting point* $x_1 \in \mathcal{C}$. *Set* $k := 1$.
Step 1. *Evaluate* $f(x_k)$ *and calculate* $g_k \in \partial f(x_k)$. *Furthermore compute*

$$\left| \begin{array}{l} f_*(k) := \min\limits_{x \in \mathcal{C}} \check{f}_k(x) \\ f^*(k) := \min\{f(x_j) : \quad 1 \leq j \leq k\} \\ x_k^* \in \operatorname{Argmin}\{f(x_j) : \quad 1 \leq j \leq k\} \end{array} \right.$$

[2] We introduce the compact set C in order to guarantee convergence of the algorithm.

Set $l_k := f_*(k) + \lambda\Delta(k)$*, where the gap* $\Delta(k)$ *is equal to* $f^*(k) - f_*(k)$*.*

Finally compute $x_{k+1} = \pi(x_k, \{x : \ x \in \mathcal{C}, \check{f}_k(x) \le l_k\})$*.*

Step 2 *(stopping test). If* $\epsilon(x_k^*) \le \epsilon$ *stop. Otherwise replace* k *by* $k+1$ *and return to step* 1.

Let S_i be the interval $[f_*(k), f^*(k)]$. Then we have

$$S_1 \supset S_2, \cdots, \qquad \text{with } |S_k| = \Delta(k).$$

In fact, by the properties of the cutting-plane function, we have

$$f_*(1) \le f_*(2) \le \ldots \le f_*(k) \le f_c^*, \qquad f^*(1) \ge f^*(2) \ge \ldots \ge f^*(k) \ge f_c^*.$$

Lemma 2.2.2 *Let* $i'' > i'$ *be such that* $\Delta(i'') \ge (1-\lambda)\Delta(i')$*. Then* $f_*(i'') \le l_{i'}$*.*

Proof. We suppose, ab absurdo, that $f_*(i'') > l_{i'}$. Consequently we have $\Delta(i'') < (1-\lambda)\Delta(i')$, which is impossible (see Fig. 2.2.1).

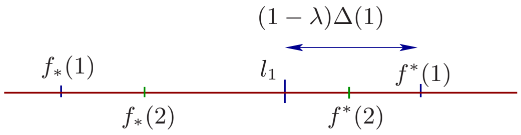


Fig. 2.2.1: $[f_*(i), f^*(i)]$.

□

Theorem 2.2.3 *For Algorithm 2.2.1. Let* $c(\lambda) = (1-\lambda)^{-2}\lambda^{-1}(2-\lambda)^{-1}$*.*

If the number of iteration N *exceeds the value* $c(\lambda)\left(\frac{LD}{\epsilon}\right)^2$*, then* $\epsilon(x_k^*) \le \epsilon$*.*

Proof. Let $I_N \triangleq \{1, \ldots, N\}$. For all $i \in I_N$, we have $\Delta(i) \ge \epsilon(x_i^*)$. We prove, by contradiction, the thesis. Suppose in fact that

$$\epsilon(x_i^*) > \epsilon \quad \forall\, i \le N,$$

where $\epsilon > 0$ is a fixed number.

We partition the index set I into the groups $I_1, \ldots, I_l$ as follows. Let $j_1 = N$. Then we define

$$I_1 \triangleq \left\{ i \le j_1 : \quad \Delta(i) \le (1-\lambda)^{-1}\Delta(j_1) \right\}.$$

Let j_2 be the largest element of I, which does not belong to I_1. Then we define

$$I_2 \overset{\triangle}{=} \left\{ i \leq j_2 : \quad \Delta(i) \leq (1-\lambda)^{-1} \Delta(j_2) \right\}.$$

Let j_p be the largest element of I, which does not belong to I_{p-1}. Then we define

$$I_p \overset{\triangle}{=} \left\{ i \leq j_p : \quad \Delta(i) \leq (1-\lambda)^{-1} \Delta(j_p) \right\}.$$

Let $l \in \{1, \dots, p\}$ and let $u(l)$ be a minimizer of $\check{f}_{j_l}(x)$ over $\mathcal{C}$, with j_l previously defined. Taking into account Lemma 2.2.2 and by virtue of the cutting-plane function properties, we have

$$\check{f}_j(u(l)) \leq \check{f}_{j_l}(u(l)) \leq l_i \quad \forall\, i, j \in I_l.$$

Consequently, $u(l) \in \mathcal{C}_i \overset{\triangle}{=} \left\{ x \in \mathcal{C} : \quad \check{f}_i(x) \leq l_i \right\}$ for all $i \in I_l$.

Let $\tau_i \overset{\triangle}{=} \|x_i - u(l)\|^2$ (see Fig. 2.2.2). For all $i \in I_l$, we have

$$\begin{aligned} \|x_i - u(l)\|^2 &= \|x_i - x_{i+1} + x_{i+1} - u(l)\|^2 = \\ &= \|x_{i+1} - x_i\|^2 + \|x_{i+1} - u(l)\|^2 - 2(u(l) - x_{i+1})^T (x_i - x_{i+1}) \\ &\geq \mathrm{dist^2(x_i \mid \mathcal{C}_i) + \tau_{i+1}} \end{aligned}$$

On the other hand, we have

$$\check{f}_i(x_i) - l_i = f(x_i) - l_i \geq f^*(i) - l_i = (1-\lambda)\Delta(i)$$

and $\check{f}_i(x_{i+1}) \leq l_i$. From above inequalities, we have

$$\check{f}_i(x_i) - \check{f}_i(x_{i+1}) \geq (1-\lambda)\Delta(i).$$

From Lipschitz property of $\check{f}_i$, it follows that

$$\mathrm{dist}(x_i \mid \mathcal{C}_i) = \|x_{i+1} - x_i\| \geq L^{-1} |\check{f}_i(x_i) - \check{f}_i(x_{i+1})| \geq L^{-1}(1-\lambda)\Delta(i).$$

Thus, for all $i \in I_l$, we have the following recurrence relation

$$\tau_{i+1} \leq \tau_i - L^{-2}(1-\lambda)^2 \Delta^2(i) \leq \tau_i - L^{-2}(1-\lambda)^2 \Delta^2(j_l).$$

Since $0 \leq \tau_i \leq D^2$, we have

$$N_l \leq D^2 L^2 (1-\lambda)^{-2} \Delta^{-2}(j_l)\ ,$$

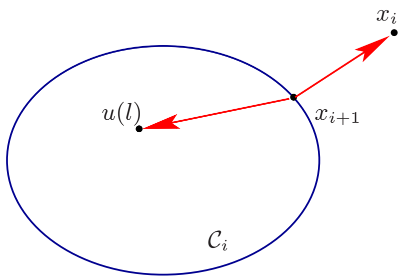


Fig. 2.2.2: $\tau_i \overset{\triangle}{=} \|x_i - u(l)\|^2$.

where N_l is the number of elements of I_l. Tacking into account the definition of I_l, we have

$$\Delta^{-1}(j_l) < \Delta^{-1}(j_1)(1-\lambda)^{l-1} < \epsilon^{-1}(1-\lambda)^{l-1}.$$

Finally $N = \sum_{l=1}^{p} N_l < D^2L^2(1-\lambda)^{-2}\epsilon^{-2}\sum_{l\geq 1}(1-\lambda)^{2(l-1)} = \left(\frac{LD}{\epsilon}\right)^2 c(\lambda)$, which is a contradiction.

□

Chapter III

Subgradient Methods and Smoothing Techniques

Introduction. We consider first the Generalized Gradient Descent method (known also as the subgradient method) introduced by N.Z. Shor in 1961 (see [38]). Then we briefly deal with the subgradient method with space dilatation in the subgradient direction.

Successively we present two systematic ways to approximate the initial nonsmooth objective function by a smooth function. In particular we describe the Moreau-Yosida regularization and a smoothing technique by Yu. Nesterov.

1 Subgradient Methods

Consider the following unconstrained minimization problem

$$(P)\begin{cases} \min\ f(x) \\ \quad x \in \mathbb{R}^n, \end{cases} \tag{1.0.1}$$

where $f \in \operatorname{Conv} \mathbb{R}^n$ is finite.

1.1 The Subgradient Method

The subgradient method is an extension of the gradient method for smooth optimization. Given a starting point x_1, the algorithm generates a sequence of points $\{x_k\}_{k=1}^{\infty}$ according to the following formula

$$x_{k+1} := x_k - t_k \frac{g_k}{\|g_k\|}, \tag{1.1.1}$$

where $t_k > 0$ is the step length and g_k is an arbitrary element of $\partial f(x_k)$. It is shown in [31] that the antisubgradient direction $-g_k$ is not necessarily a descent direction, as it is in the smooth optimization.

Let $x_k \notin \operatorname*{Argmin}_{\mathbb{R}^n} f$, $x^* \in \operatorname*{Argmin}_{\mathbb{R}^n} f$. From subgradient inequality, we have

$$\begin{aligned} \|x_{k+1} - x^*\|^2 &= \|x_k - x^*\|^2 + t_k^2 - 2t_k \frac{g_k^T}{\|g_k\|}(x_k - x^*) \\ &< \|x_k - x^*\|^2 \end{aligned}$$

for $0 < t_k < -2\dfrac{g_k^T}{\|g_k\|}(x^* - x_k)$. Therefore, for small step sizes t_k, even if descent is not achieved, the distance between the current iterate and x^* decreases. This observation is the basis for all subgradients methods.

Theorem 1.1.1 [39] *Let $f \in \operatorname{Conv} \mathbb{R}^n$ be finite such that $\operatorname*{Argmin}_{\mathbb{R}^n} f$ is bounded. Let $\{t_k\}_{k=1}^{\infty}$ be a sequence of positive real numbers satisfying the following relations:*

$$\lim_{k\to\infty} t_k = 0, \qquad \sum_{k=1}^{\infty} t_k = +\infty \ .$$

Then the sequence $\{x_k\}_{k=1}^{\infty}$ obtained by (1.1.1), for any starting point x_1, satisfies one of the following properties: either there exists $\bar{k}$ such that $x_{\bar{k}} \in \operatorname*{Argmin}_{\mathbb{R}^n} f$ or

$$\lim_{k\to\infty} \min\{\|x_k - x\| : \quad x \in \operatorname*{Argmin}_{\mathbb{R}^n} f\} = 0, \quad \lim_{k\to\infty} f(x_k) = \min_{\mathbb{R}^n} f$$

This method was extended to the convex programming problems in Hilbert spaces by B.T. Polyak [32].

1.2 The Subgradient Method with Space Dilatation in the Subgradient Direction

When the upper bounds of the angles between the antigradients $-g_k \in \partial f(x_k)$ and the directions[1] $x^* - x_k$ are equal to $\frac{\pi}{2}$, then the convergence of the subgradient algorithm is very slow. To cope with this problem, at each iteration k it is introduced a linear operator changing the metric of the space (see [38]).

Let $\xi \in \mathbb{R}^n$ be a vector such that $\|\xi\| = 1$. Then any point $x \in \mathbb{R}^n$ may be represented as follows

$$x = \gamma_\xi(x)\xi + d_\xi(x) \tag{1.2.1}$$

where $\xi^T d_\xi(x) = 0$.

Consequently, we have $\gamma_\xi(x) = x^T\xi$ and $d_\xi(x) = x - \xi\xi^T x$.

Definition 1.2.1 [38] *For a given number $\alpha \geq 0$ and a vector $\xi \in \mathbb{R}^n$, $\|\xi\| = 1$. The operator in vector form $R_\alpha(\xi)$,*

$$\begin{aligned} x \mapsto R_\alpha(\xi)x &= [I_n + (\alpha - 1)\xi\xi^T]x \\ &= \alpha\gamma_\xi(x)\xi + d_\xi(x). \end{aligned} \tag{1.2.2}$$

is called an operator of space dilatation along direction ξ with coefficient α.

Now we present the subgradient algorithm with space dilatation in the subgradient direction. Given a nonsingular matrix B_k, let $y \mapsto \varphi_k(y) \overset{\triangle}{=} f(B_k y)$ be a function obtained from f taking into account the linear transformation $y = B_k^{-1}x$.

Algorithm 1.2.1 [39]

Step 0. *Choose a starting point x_1 and put $B_1 := I_n$. Set $k := 1$.*

Step 1. *Calculate an arbitrary subgradient $g_k \in \partial f(x_k)$.*

Step 2. *Determine $\hat{g}_k := B_k^T g_k$, where $\hat{g}_k$ is an particular subgradient of the function φ_k, defined in the "dilated" space, at $y_k = B_k^{-1}x_k$.*

Step 3. *Put $\xi_k := \frac{\hat{g}_k}{\|\hat{g}_k\|}$. Find a scalar t_k and, successively, put $x_{k+1} := x_k - t_k B_k \xi_k$.*

Step 3. *Select a coefficient $\alpha_k > 1$ of space dilatation and update the the matrix of space transformation*

$$B_{k+1} := B_k R_{\frac{1}{\alpha_k}}(\xi_k). \tag{1.2.3}$$

[1] $x^* \in \operatorname{Argmin}_{\mathbb{R}^n} f$.

From the point of view of dilated space, it results that

$$y_{k+1} = R_{\alpha_k}(\xi_k)\,[y_k - t_k\xi_k]\ .$$

Step 4. *Replace* k *by* $k+1$ *and go to step* 2.

As regards convergence, the coefficient α_k of space dilatation plays the main role. N.Z. Shor proved that the Ellipsoid method, introduced by D.B. Judin and A.S. Nemirovskij, is a special case of algorithm with space dilatation in the subgradient direction [39].

2 Smoothing Techniques

Now we present two smoothing techniques.

2.1 Moreau-Yosida Regularization

Consider the following unconstrained minimization problem

$$(P)\begin{cases} \min\ f(x) \\ \quad x \in \mathbb{R}^n, \end{cases} \tag{2.1.1}$$

where $f \in \overline{\text{Conv}}\ \mathbb{R}^{\text{n}}$.

Let Q be a positive definite matrix. Then the function

$$x \mapsto f_Q(x) \triangleq \min_{y\in\mathbb{R}^n}\left\{f(y) + \frac{1}{2}(y-x)^T Q(y-x)\right\} : \mathbb{R}^n \to \mathbb{R}\cup\{+\infty\} \tag{2.1.2}$$

is called the "Moreau-Yosida regularization" of f associated with Q.

Lemma 2.1.1 [16] *The minimization problem involved in (2.1.2) is well-posed and has unique solution, characterized as the unique point* $y \in \mathbb{R}^n$ *satisfying*

$$Q(x-y) \in \partial f(y).$$

The unique solution of the minimization problem in (2.1.2), denoted as $p_Q(x)$, is called the "proximal point" of x associated with f and Q.

Consequently, we have

$$p_Q(x) = x - Q^{-1}g_Q(x)\ ,$$

where $g_Q(x)$ is the particular subgradient of f at $p_Q(x)$ defined via Lemma 2.1.1. We remark that $g_Q(x)$ must not be confused with an arbitrary subgradient of f at $p_Q(x)$.

Theorem 2.1.2 [16] *The function f_Q in (2.1.2) is finite everywhere, convex and differentiable; its gradient is*

$$\nabla f_Q(x) = g_Q(x) = Q(x - p_Q(x))$$

and

$$\|\nabla f_Q(x) - \nabla f_Q(y)\| \leq \lambda_{\max}(Q)\|x - y\| \quad \forall\ x, y \in \mathbb{R}^n \ ,$$

where $\lambda_{\max}(Q)$ is the largest eigenvalue of Q.

Consequently, the function f_Q is continuously differentiable (or smooth). It is an approximation of f as well; in fact we have the following result.

Proposition 2.1.3 [16] *Let $\lambda_{\min}(Q)$ be the smallest eigenvalue of Q. As $\lambda_{\min}(Q) \to +\infty$, $f_Q(x)$ tends to $f(x)$ for every $x \in \mathbb{R}^n$ and $p_Q(x)$ tends to x for all $x \in \operatorname{dom} f$.*

The basic properties of the Moreau-Yosida regularization can be summarized by the following theorem.

Theorem 2.1.4 [16] *Let f_Q be the Moreau-Yosida regularization of f. Minimizing f and f_Q are equivalent problems, in the sense that*

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

and the following statements are equivalent:

(i) *x minimizes f;*

(ii) *$p_Q(x) = x$;*

(iii) *$g_Q(x) = 0$;*

(iv) *x minimizes f_Q;*

(v) *$f(p_Q(x)) = f(x)$;*

(vi) *$f_Q(x) = f(x)$.*

Consequently, (ii) seems to suggest to design an algorithm finding a fixed point of the mapping $x \mapsto p_Q(x)$. According to this idea, the following algorithm aims to solve the problem (P).

Algorithm 2.1.5 (Proximal Point algorithm)[16]
Step 0. *Select an initial point $x_1 \in \mathbb{R}^n$. Choose an initial positive definite matrix Q_1. Set $k := 1$.*
Step 1. *Calculate x_{k+1} by finding the proximal point of x associated with Q_k, i.e.*

$$x_{k+1} := p_{Q_k}(x_k) = \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{1}{2}(y - x_k)^T Q_k (y - x_k) \right\} .$$

Step 2. *If $x_{k+1} := x_k$ stop.*
Step 3. *Choose a new positive definite matrix Q_{k+1}. Replace k by $k + 1$ and return to step* 1.

Algorithm 2.1.5 is just an abstract scheme, since is not clear how to compute x_{k+1} at step 1. There exist some improvements of this method (see, e.g., [16]).

2.2 Nesterov's Technique

Consider the convex problem (P) defined in §II.1. Let $\epsilon > 0$ be the requested absolute accuracy in the solution of (P), that is it is required to find $x \in \mathcal{C}$ such that

$$f(x) - \min_{\mathcal{C}} f \leq \epsilon \ .$$

The lower complexity bound of the standard subgradient method (see, e.g., [27]) is [2] of the order $O\left(\frac{1}{\epsilon^2}\right)$. It was also proved in [26] that it is not possible to improve the efficiency estimate uniformly in the dimension of the space of variables, that is the theoretical lower complexity of (P) is of the order $O\left(\frac{1}{\epsilon^2}\right)$. In fact the following problem is difficult for all numerical algorithm schemes:

$$\begin{cases} \min\limits_{x \in \mathbb{R}^n} \max\limits_{1 \leq i \leq n} x_i \\ \quad \sum_{i=1}^n x_i^2 \leq 1 \ . \end{cases}$$

[2] $O(t)$: there exists $C > 0$ such that $|O(t)| \leq C|t|$; in this case it is the bound on the number of iterations.

We restrict our attention to functions endowed with a special explicit structure, i.e. f is a function of the form

$$x \mapsto f(x) = \hat{f}(x) + \max_{\lambda \in \bar{\mathcal{C}}} \left\{ x^T Q \lambda - \hat{\phi}(\lambda) \right\} \ ,$$

where $\bar{\mathcal{C}} \in \mathbb{R}^m$ is a compact convex set, $\hat{f}$ is continuously differentiable and convex on $\mathcal{C}$, $\hat{\phi}$ is continuously differentiable and convex on $\bar{\mathcal{C}}$ and Q is a matrix of appropriate dimension.

For the class of functions above considered, the problem (P) can be solved with efficiency estimate of the order $O\left(\frac{1}{\epsilon}\right)$.

Let $d_2 : \bar{\mathcal{C}} \to \mathbb{R}$ be convex and continuously differentiable such that

$$d_2(\lambda) \geq \frac{1}{2}\sigma_2 \|\lambda - \lambda_0\|^2 + (\lambda - \lambda_0)^T \nabla d_2(\lambda_0) + d_2(\lambda_0) \quad \forall\ \lambda, \lambda_0 \in \bar{\mathcal{C}} \ ,$$

for some $\sigma_2 > 0$. Then d_2 is called a prox-function of $\bar{\mathcal{C}}$ with parameter σ_2. Let $\lambda_0 = \operatorname*{argmin}_{\bar{\mathcal{C}}} d_2$ and assume without loss of generality that $d_2(\lambda_0) = 0$. Consequently, we have

$$d_2(\lambda) \geq \frac{1}{2}\sigma_2 \|\lambda - \lambda_0\|^2 \quad \forall\ \lambda \in \bar{\mathcal{C}} \ . \tag{2.2.1}$$

Theorem 2.2.1 [29] *Let d_2 be a prox-function of $\bar{\mathcal{C}}$ of the form (2.2.1) and let μ_2 be a positive number. Then the function*

$$f_{\mu_2}(x) \overset{\triangle}{=} \max_{\lambda \in \bar{\mathcal{C}}} \left\{ x^T Q \lambda - \hat{\phi}(\lambda) - \mu_2 d_2(\lambda) \right\} \tag{2.2.2}$$

is convex and continuously differentiable on $\mathbb{R}^n$ and its gradient at a point x is

$$\nabla f_{\mu_2}(x) = Q \lambda_{\mu_2},$$

where λ_{μ_2} is the unique solution of the optimization problem involved in (2.2.2).

Furthermore, the function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitzian with constant

$$\frac{1}{\mu_2 \sigma_2} \|A\|^2 \ ,$$

where $\|A\| \overset{\triangle}{=} \max_{\|x\|=1} \|Q^T x\|$.

Let $D_2 \overset{\triangle}{=} \max_{\bar{\mathcal{C}}} d_2$ and $\bar{f}_{\mu_2}(x) \overset{\triangle}{=} \hat{f}(x) + f_{\mu_2}(x)$. Consequently, we have

$$\bar{f}_{\mu_2}(x) \leq f(x) \leq \bar{f}_{\mu_2}(x) + \mu_2 D_2 \quad \forall\; x \in \mathbb{R}^n. \tag{2.2.3}$$

Obviously, the function $\bar{f}_{\mu_2}$, for $\mu_2 > 0$, is convex and continuously differentiable on $\mathcal{C}$. Thus, $\bar{f}_{\mu_2}$ can be seen as a uniform smooth approximation of the function f.

2.3 Excessive Gap Technique

By considering the particular form of the objective function f, discussed in §2.2, we can write the problem (P) in an "adjoint form":

$$(D) \begin{cases} \max \phi(\lambda) \\ \quad \lambda \in \bar{\mathcal{C}}, \end{cases}$$

where $\phi(\lambda) = -\hat{\phi}(\lambda) + \min_{x \in \mathcal{C}} \left\{ x^T Q \lambda + \hat{f}(x) \right\}$.

It is clear that

$$\phi(\lambda) \leq f(x) \qquad \forall \lambda \in \bar{\mathcal{C}},\; \forall x \in \mathcal{C}\ .$$

From Sion-Kakutani Theorem, taking into account that $L(x, \lambda) = -\hat{\phi}(\lambda) + x^T Q \lambda + \hat{f}(x)$ is continuous and convex in $x \in \mathcal{C}$ for every fixed $\lambda \in \bar{\mathcal{C}}$ and is continuous and concave in $\lambda \in \bar{\mathcal{C}}$ for every fixed $x \in \mathcal{C}$ and that $\mathcal{C}$ and $\bar{\mathcal{C}}$ are compact convex sets, it follows (see, e.g., [2]) that

$$\max_{\bar{\mathcal{C}}} \phi = \min_{\mathcal{C}} f\ .$$

Similarly, let d_1 be a prox-function of $\mathcal{C}$ of the form (2.2.1) with parameter σ_1 and let μ_1 be a positive number. Then the function

$$\bar{\phi}_{\mu_1}(\lambda) \overset{\triangle}{=} -\hat{\phi}(\lambda) + \min_{x \in \mathcal{C}} \left\{ x^T Q \lambda + \hat{f}(x) + \mu_1 d_1(x) \right\}$$

is concave and continuously differentiable on $\bar{\mathcal{C}}$. Of course, we have

$$\bar{\phi}_{\mu_1}(\lambda) - \mu_1 D_1 \leq \phi(\lambda) \leq \bar{\phi}_{\mu_1}(\lambda) \quad \forall\; \lambda \in \mathbb{R}^m\ , \tag{2.3.1}$$

where $D_1 \overset{\triangle}{=} \max_{\mathcal{C}} d_1$.

Lemma 2.3.1 *Let $x \in \mathcal{C}$ and $\lambda \in \bar{\mathcal{C}}$. If the "excessive gap condition"*

$$\bar{f}_{\mu_2}(x) \leq \bar{\phi}_{\mu_1}(\lambda)$$

is satisfied, then

$$0 \leq f(x) - \phi(\lambda) \leq \mu_1 D_1 + \mu_2 D_2 \ .$$

Proof. From (2.2.3) and (2.3.1), we have

$$f(x) - \mu_2 D_2 \leq \bar{f}_{\mu_2}(x) \leq \bar{\phi}_{\mu_1}(\lambda) \leq \phi(\lambda) + \mu_1 D_1.$$

Consequently, the thesis follows.

□

Now we require that the optimization problems involved in the definitions of f and ϕ are solved in a closed form. We assume also that the structures of the objects $\hat{f}$, $\hat{\phi}$, $\mathcal{C}$ and $\bar{\mathcal{C}}$ are simple enough and that the functions $\hat{f}$ and $\hat{\phi}$ have Lipschitzian gradients. Under these hypotheses is described in [28] an algorithm which generates a sequences of pairs[3] $\{x_k, \lambda_k\}_{k\in\mathbb{N}}$ satisfying the excessive gap condition and such that

$$f(x_k) - \phi(\lambda_k) \leq \frac{4\|A\|}{k+1}\sqrt{\frac{D_1 D_2}{\sigma_1\sigma_2}} \qquad \forall\ k \in \mathbb{N}.$$

Letting

$$\epsilon \leq \frac{4\|A\|}{k+1}\sqrt{\frac{D_1 D_2}{\sigma_1\sigma_2}} \ ,$$

the rate of convergence is $\epsilon = O\left(\frac{1}{k}\right)$.

[3] Either the parameter μ_1 or μ_2 is decremented at each iteration k.

Nonsmooth Nonconvex Optimization

Chapter IV

Some Elements of Nonsmooth Analysis

Introduction. The notion of the derivative was introduced by G.W. Leibnitz in Nova Methodus (1684) (see, e.g., [1]). Since 1684 the differential calculus has been more and more enriched by new definitions of differentiability. We analyze the Fréchet, the Gâuteaux and the directional differentiability and we discuss some results on both the theory of Clarke gradient and Goldstein ϵ-subdifferential. Then we present two mean value theorems, one for directionally differentiable functions and the other for locally Lipschitzian functions, and we define the semismoothness and weak semismoothness. Finally we report a proof of Rademacher Theorem.

1 Differentiability

The derivative of a real-valued function on $\mathbb{R}$ gives a measure of how the function changes when its argument changes.

Definition 1.0.1 (Derivative)[40] *Let f be a real-valued function defined on an interval (y, z) and let $x \in (y, z)$. If the limit*

$$f'(x) = \lim_{d \to 0} \frac{f(x + d) - f(x)}{d} \tag{1.0.1}$$

exists, then f is differentiable at x. The value of the limit (1.0.1) is called the derivative of f at x.

If we consider functions of several real variables instead of functions of just a real variable, since d is a vector in $\mathbb{R}^n$, then Definition 1.0.1 is without meaning. On the other hand, Definition 1.0.1 suggests how to approximate f near x; in particular if f is differentiable at x, then there exists a linear function $d \to f'(x)d$ on $\mathbb{R}$ taking values in $\mathbb{R}$ which approximates the change $f(x+d) - f(x)$ in f up to a remainder[1] which is $o(|d|)$:

$$|f(x+d) - f(x) - f'(x)d| \leq o(|d|) \ .$$

Using last formulation of the derivative, it is possible to deal with the differentiability for real-valued functions of several real variables.

1.1 Fréchet Differentiability

A real-valued function on $\mathbb{R}^n$ is Fréchet differentiable, if it is differentiable in the classical sense.

Definition 1.1.1 (Fréchet differentiability)[2] *Let f be a real-valued function defined on an open set $A \subset \mathbb{R}^n$ and let x be a point of A. Then f is said to be Fréchet differentiable (or just differentiable) at x, if there exists a linear function $L_x(d) : \mathbb{R}^n \to \mathbb{R}$, called derivative of f at x, which approximates the change $f(x+d) - f(x)$ in f up to a remainder which is $o(\|d\|)$:*

$$|f(x+d) - f(x) - L_x(d)| \leq o(\|d\|). \tag{1.1.1}$$

Equivalently: f is called Fréchet differentiable at x, if there exists a linear function $d \mapsto L_x(d)$ on $\mathbb{R}^n$ taking values in $\mathbb{R}$ such that for every $\epsilon > 0$ there exists $\delta_\epsilon > 0$ satisfying the relation

$$\|d\| \leq \delta_\epsilon \ \Rightarrow \ |f(x+d) - f(x) - L_x(d)| \leq \epsilon\|d\|. \tag{1.1.2}$$

Moreover if f is differentiable at each point of A, then f is said to be Fréchet differentiable (or just differentiable) on A.

[1] $o(|d|)$ indicates all functions of d which vanish at $d = 0$ and are such that the ratio $\frac{o(d)}{|d|}$ approaches zero as $|d| \to 0$.

The linear function L_x satisfying (1.1.1) is unique. This basic result is established by following lemma.

Lemma 1.1.1 (Uniqueness of the derivative) [2] *Let f be a real-valued function defined on an open set $A \subset \mathbb{R}^n$ and differentiable at a point $x \in A$. Then the derivative $L_x(d) : \mathbb{R}^n \to \mathbb{R}$ participating in Definition 1.1.1 is*

$$L_x(d) = \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t} \ . \tag{1.1.3}$$

In particular, L_x is uniquely defined by f and x.

Proof. Choose any $v \in \mathbb{R}^n$. Substituting $d = tv$ whit $t > 0$ into (1.1.1) and taking into account that L_x is linear, we have

$$\left| \frac{f(x + tv) - f(x)}{t} - L_x(v) \right| \leq \frac{o(t\|v\|)}{t}.$$

The thesis follows by passing to limit as $t \downarrow 0$.

□

Taking into account that L_x is a linear function, then the derivative can be represented as

$$L_x(d) = \nabla f(x)^T d,$$

where $\nabla f(x)$ is a vector and it is called the gradient of f at x.

1.2 Directional Differentiability

Given two vectors $x, d \in \mathbb{R}^n$ and a real-valued function f on $\mathbb{R}^n$, we define the unidimensional function $\phi(t) \overset{\triangle}{=} f(x + td)$. The directional derivative of f at x along d is defined as the right-hand-side derivative of ϕ at 0. It is gives the approximate change in ϕ for a small step $t > 0$.

Definition 1.2.1 (Directional derivative)[2] *Let f be a real-valued function defined on an open set $A \subset \mathbb{R}^n$. Let x be a point of A and let d be a vector in $\mathbb{R}^n$. If the limit*

$$\lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t} \tag{1.2.1}$$

exists, it is called the directional derivative of f at x along the direction d and is denoted by $f'(x,d)$.

The importance of the directional derivative is based on the fact that it is easy to conceive. Of course if we restrict ourselves to the unidimensional case, all the quantities at stake assume a meaning easier to guess.

Definition 1.2.2 (Directional differentiability)[7] *Let f be a real-valued function defined on an open set $A \subset \mathbb{R}^n$ and let x be a point of A. The function f is said directionally differentiable at x if the limit (1.2.1) exists for every $d \in \mathbb{R}^n$. Finally f is said directionally differentiable on A if it is directionally differentiable at every point of A.*

It follows that if f is differentiable at x, then the value of the derivative L_x at a point d coincides with the directional derivative of f at x in the direction d.

Remark 1.2.1 *If the function f participating in Definition 1.2.2 is directionally differentiable at a point x and $d \mapsto f'(x,d)$ is a linear function of d, then $f'(x,d) = -f'(x,-d)$ and consequently*

$$\lim_{t\to 0}\frac{f(x+td)-f(x)}{t} = \lim_{t\downarrow 0}\frac{f(x+td)-f(x)}{t} = \lim_{t\uparrow 0}\frac{f(x+td)-f(x)}{t} \ .$$

1.3 Gâuteaux Differentiability and Examples

Definition 1.3.1 (Gâuteaux differentiability)[25] *Let f be a real-valued function defined on an open set $A \subset \mathbb{R}^n$ and let x be a point of A. Then f is said to be Gâuteaux differentiable at x, if there exists a linear function $d \mapsto f^*(x,d)$ on $\mathbb{R}^n$ taking values in $\mathbb{R}$ such that*

$$|f(x+td)-f(x)-tf^*(x,d)| \leq o(|t|) \ . \tag{1.3.1}$$

Equivalently: f is called Gâuteaux differentiable at x, if there exists a linear func-

tion $d \mapsto f^*(x,d)$ on $\mathbb{R}^n$ taking values in $\mathbb{R}$ such that

$$\begin{gathered} \forall\ d > 0,\ \forall\ \epsilon > 0\ \ \exists\ \delta_{\epsilon,d} > 0\ s.t. \\ |t| \leq \delta_{\epsilon,d}\ \Rightarrow\ |f(x+td) - f(x) - tf^*(x,d)| \leq \epsilon|t| \end{gathered} \tag{1.3.2}$$

or simply there exists the limit

$$f^*(x,d) \triangleq \lim_{t\to 0} \frac{f(x+td) - f(x)}{t}, \quad \forall\ d \in \mathbb{R}^n\ . \tag{1.3.3}$$

Moreover if f is Gâuteaux differentiable at each point of A, then f is said to be Gâuteaux differentiable on A.

In other words, f is called Gâuteaux differentiable at $x \in A$, if it is directionally differentiable at x and the directional derivative[2] $f'(x,d)$, viewed as function of d, is linear.

It remains to understand what is the difference between the Fréchet differentiability and Gâuteaux differentiability; it can be viewed by analyzing the role of δ in (1.1.2) and (1.3.2). We observe that in the last formula δ depends on both ϵ and d, whereas in (1.1.2) only on ϵ.

Now we examine more formally the relationship between the Fréchet differentiability and Gâuteaux differentiability.

Theorem 1.3.1 *Let f be a real-valued function defined on an open set $A \subset \mathbb{R}^n$ and let x be a point of A. Then f is Fréchet differentiable at x if and only if it is Gâuteaux differentiable at x and the limit (1.3.3) is uniform on the unit sphere $S^{(n)}$.*

Proof.

($\Rightarrow$) Let $v \in \mathbb{R}^n$ and take $d = tv$, for $t > 0$. It follows from (1.1.1) that

$$0 = \lim_{\|d\|\to 0} \frac{f(x+d) - f(x) - L_x(d)}{\|d\|} = \lim_{t\downarrow 0} \frac{f(x+tv) - f(x) - tf'(x,v)}{t\|v\|}$$

Thus f is Gâuteaux differentiable at x. To prove that the limit (1.3.3) is uniform on $S^{(n)}$, take any $v \in S^{(n)}$. Applying (1.1.2) with $d = tv$, we have that for every $\epsilon > 0$ there exists $\delta_\epsilon > 0$ satisfying the relation

$$|t| \leq \delta_\epsilon\ \Rightarrow\ |f(x+tv) - f(x) - tf'(x,v)| \leq \epsilon|t|\ . \tag{1.3.4}$$

[2] By considering Remark 1.2.1 $f'(x,d)$ coincides with $f^*(x,d)$

($\Leftarrow$) Taking into account that any vector $v \in \mathbb{R}^n$ can be represented as $v = td$, for some $d \in S^{(n)}$ and $t > 0$, and the limit (1.3.3) is uniform on $S^{(n)}$, we derive from (1.3.2) that for every $\epsilon > 0$ there exists $\delta_\epsilon > 0$ such that

$$\|v\| \leq \delta_\epsilon \;\Rightarrow\; |f(x+v) - f(x) - f'(x,v)| \leq \epsilon \|v\|.$$

□

The following example shows that the Fréchet differentiability not always coincides with the Gâuteaux differentiability.

Example 1.3.1 (Apple function)[6] *Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$, displayed in Fig. 1.3.1,*

$$f(x) = \begin{cases} 1 & x \in \mathcal{C} \\ 0 & x \notin \mathcal{C} \ . \end{cases}$$

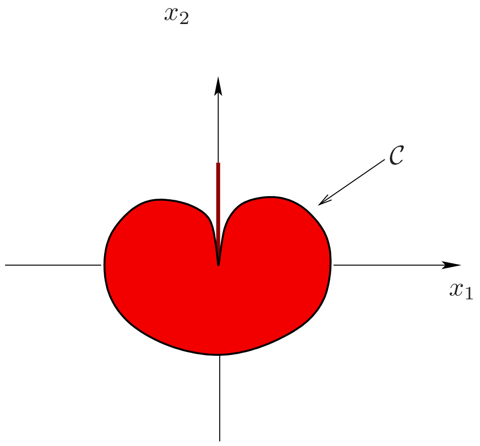


Fig. 1.3.1: Apple function.

Take $\bar{x} = (0,0)^T$ and remark that the stalk of the "apple" ($\mathcal{C}$), shown in Fig. 1.3.1, is a vertical tangent at $\bar{x}$ to the curve forming the boundary of the "apple".

(i) *For any $d \in \mathbb{R}$, we have*

$$f'(\bar{x}, d) = \lim_{t \downarrow 0} \frac{f(\bar{x} + td) - f(x)}{t} = 0$$

Consequently f is Gâuteaux differentiable at $\bar{x}$.

(ii) *Let $\epsilon = \frac{1}{2}$. For every $\delta > 0$ there exists, for some d of the unit sphere $S^{(2)}$, a point $\bar{x} + td \notin S^{(2)}$, with $0 < t < \min\{\delta, 2\}$, satisfying the inequality*

$$\left| \frac{f(\bar{x} + td) - f(\bar{x}) - tf'(\bar{x}, d)}{t} \right| = \frac{1}{t} > \frac{1}{2},$$

which contradicts (1.3.2). It follows that f is not Fréchet differentiable at $\bar{x}$.

The following simple example proves that the directional differentiability does not imply the Gâuteaux differentiability.

Example 1.3.2 *Consider the real-valued function f of a real variable, shown in Fig. 1.3.2,*

$$f(x) = |x| \ .$$

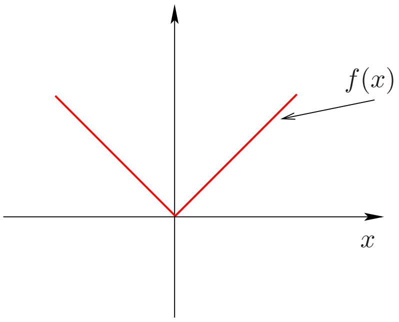


Fig. 1.3.2: $f(x) = |x|$.

Let $\bar{x} = 0$. For any $d \in \mathbb{R}$, from (1.2.1), we have

$$f'(\bar{x}, d) = \lim_{t \downarrow 0} \frac{f(\bar{x} + td) - f(\bar{x})}{t} = |d| \ .$$

Therefore f is directionally differentiable at $\bar{x}$, but the function $d \mapsto f'(\bar{x}, d)$ is not linear. Hence f is not Gâuteaux differentiable at $\bar{x}$.

1.4 Continuity and Differentiability

Now we analyze the relationship between continuity and differentiability.

Theorem 1.4.1 *Let f be a real-valued function defined on an open set $A \subset \mathbb{R}^n$ and let x be a point of A. If f is Fréchet differentiable at x, then f is continuous at x.*

Proof. Choose any $d \in \mathbb{R}^n$. Taking into account that the derivative $L_x(d)$ is linear in d, it follows from (1.1.1) that

$$\lim_{d \to 0} f(x+d) - f(x) = \lim_{d \to 0} \frac{f(x+d) - f(x)}{\|d\|} \|d\| = \lim_{d \to 0} L_x \left(\frac{d}{\|d\|} \right) \|d\| = 0 \ .$$

□

Remark 1.4.1 *Continuity is not implied by Gâuteaux differentiability; in fact the function considered in Example 1.3.1 is Gâuteaux differentiable and not continuous at $\bar{x}$.*

In smooth optimization, we assume that the functions involved are continuously differentiable on $\mathbb{R}^n$.

Definition 1.4.1 (Continuous differentiability)[2] *Let f be a real-valued function defined on an open set $A \subset \mathbb{R}^n$ and Fréchet differentiable at $x \in A$. If the gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is continuous at x, then f is said to be continuously differentiable at x. If f is continuously differentiable at each $x \in A$, then f is called continuously differentiable (or smooth) on A.*

The set of continuously differentiable functions on $\mathbb{R}^n$ taking values in the real axis is denoted by C^1.

The continuous differentiability at a point $x \in \mathbb{R}^n$ is not a consequence of the Fréchet differentiability as shown in Example 1.4.1.

Example 1.4.1 *Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$*

$$f(x_1, x_2) = \begin{cases} x_1^2 \sin \frac{1}{x_1} + x_2 & x_1 \neq 0 \\ x_2 & x_1 = 0 \ . \end{cases}$$

Let $d = (d_1, d_2)^T$ and $x = (x_1,\ x_2)^T$. Take $\bar{x} = (0, 0)^T$.

(i) *Compute the directional derivative of f at $\bar{x}$:*

$$f'(\bar{x},d) = \lim_{t\to 0} \frac{f(\bar{x}+td)-f(\bar{x})}{t} = \lim_{t\to 0} \frac{t^2 d_1^2 \sin\left(\dfrac{1}{td_1}\right) + td_2}{t} = d_2 \ .$$

Consequently f is Gâuteaux differentiable at $\bar{x}$. For every d in the unit sphere $S^{(2)}$, for every $\epsilon > 0$ there exists $\delta = \epsilon$ satisfying

$$|t| \le \delta \Rightarrow \left| \frac{f(\bar{x}+td)-f(\bar{x})-tf'(\bar{x},d)}{t} \right| = \left| td_1^2 \sin\frac{1}{td_1} \right| \le |t| \le \epsilon \ .$$

By Theorem 1.3.1, f is Fréchet differentiable at $\bar{x}$ and $\nabla f(\bar{x}) = (0,1)^T$.

(ii) *Let $\{x^k\}_{k\in\mathbb{N}}$ such that $x^k = \left(\frac{1}{2k\pi}, \frac{1}{k}\right)^T$. Then $\lim\limits_{k\to\infty} x^k = \bar{x}$ and*

$$\lim_{k\to\infty} \nabla f(x^k) = \lim_{k\to\infty} \begin{pmatrix} \frac{1}{k\pi}\sin(2k\pi) - \cos(2k\pi) \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \ne \nabla f(\bar{x}) \ .$$

Therefore f is not continuously differentiable at $\bar{x}$.

1.5 Lipschitzianity and Differentiability

Now we illustrate the differential properties of Lipschitzian functions.

Definition 1.5.1 (Directional Dini derivatives)[25] *Let f be a real-valued function defined on an open set $A \subset \mathbb{R}^n$. Let x be a point of A and let d be a vector in $\mathbb{R}^n$. Then the upper and lower directional Dini derivative of f at x along the direction d are given respectively by*

$$\begin{aligned} \overline{f}'(x,d) &\triangleq \limsup_{t\to 0} \frac{f(x+td)-f(x)}{t} \\ &\triangleq \lim_{\tau\downarrow 0} \overline{g}^x_\tau(d) \triangleq \lim_{\tau\downarrow 0} \sup_{t\in(-\tau,\tau)\setminus\{0\}} \left\{ \frac{f(x+td)-f(x)}{t} \right\} \end{aligned} \tag{1.5.1a}$$

$$\begin{aligned} \underline{f}'(x,d) &\triangleq \liminf_{t\to 0} \frac{f(x+td)-f(x)}{t} \\ &\triangleq \lim_{\tau\downarrow 0} \underline{g}^x_\tau(d) \triangleq \lim_{\tau\downarrow 0} \inf_{t\in(-\tau,\tau)\setminus\{0\}} \left\{ \frac{f(x+td)-f(x)}{t} \right\} \end{aligned} \tag{1.5.1b}$$

It is clear from the properties of the limits (see, e.g., [36]) that the directional Dini derivatives always exist as extended real numbers.

The following proposition summarizes the relationship between Fréchet differentiability and Gâuteaux differentiability for Lipschitzian functions.

Proposition 1.5.1 *Let f be a real-valued function defined and Lipschitzian on an open set $A \subset \mathbb{R}^n$ and let x be a point of A. Then f is Fréchet differentiable at x if and only if it is Gâuteaux differentiable at x.*

Proof.

($\Rightarrow$) See Theorem 1.3.1.

($\Leftarrow$) Under the hypothesis that the function f is Lipschitzian it is known (see, e.g., [25]) that the limits (1.5.1) are uniform on the unit sphere $S^{(n)}$.

Thus, taking into account that the directional derivative coincides with the directional Dini derivatives, it follows that for every $\epsilon > 0$ there exists a $\delta_\epsilon > 0$ such that

$$|t| < \delta_\epsilon \ \Rightarrow -\epsilon + f'(x,d) < \frac{f(x+td) - f(x)}{t} < f'(x,d) + \epsilon.$$

The thesis follows by using Theorem 1.3.1.

□

2 Subdifferentials of Nonconvex Functions

Nonsmooth analysis studies have greatly benefited from properties of Clarke gradient which is a systematic extension of the subdifferential of convex functions to the family of locally Lipschitzian functions.

2.1 Clarke Gradient

The Clarke gradient is developed for real-valued locally Lipschitzian functions defined on a Banach space S (see [5]); we restrict ourselves to the finite-dimensional case, that is $S = \mathbb{R}^n$.

Definition 2.1.1 (Clarke derivative)[5] *Let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitzian near $x \in \mathbb{R}^n$. Then the Clarke derivative (or generalized directional derivative) of f at x along a direction d, denoted as $f^o(x,d)$, is given by*

$$\limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y+td)-f(y)}{t}. \tag{2.1.1}$$

Moreover f is said to be "regular" at x if f is directionally differentiable at x and

$$f'(x,d) = f^o(x,d) \quad \forall\ d \in \mathbb{R}^n.$$

The Clarke derivative was introduced by F.H. Clarke in the Seventies [5].

Definition 2.1.2 (Clarke gradient)[5] *Let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitzian near $x \in \mathbb{R}^n$. Then the Clarke gradient (or generalized gradient) of f at x, denoted as $\partial f(x)$, is given by*

$$\{g \in \mathbb{R}^n : \quad g^T d \le f^o(x,d) \ \ \forall\ d \in \mathbb{R}^n\}$$

We remark that Clarke gradient is not a generalization of the concept of Fréchet differentiability as is proved by following example.

Example 2.1.1 *Consider the function $f : \mathbb{R} \to \mathbb{R}$, displayed in Fig. 2.1.1,*

$$f(x) = \begin{cases} x^2 \sin \dfrac{1}{x} & x \neq 0 \\ 0 & x = 0\ . \end{cases}$$

Let $\bar{x} = 0$. It is clear that f is Lipschitzian near $\bar{x}$. We have $f'(\bar{x}) = 0$, while $\partial f(\bar{x}) = co\{-1, 1\}$.

Proposition 2.1.1 [5] *Let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitzian near $x \in \mathbb{R}^n$ and let d be a vector in $\mathbb{R}^n$. Then the Clarke derivative is finite and sublinear*[3] *and*

$$f^o(x,d) = \max_{g \in \partial f(x)} g^T d\ .$$

[3] see Definition I.2.1.1

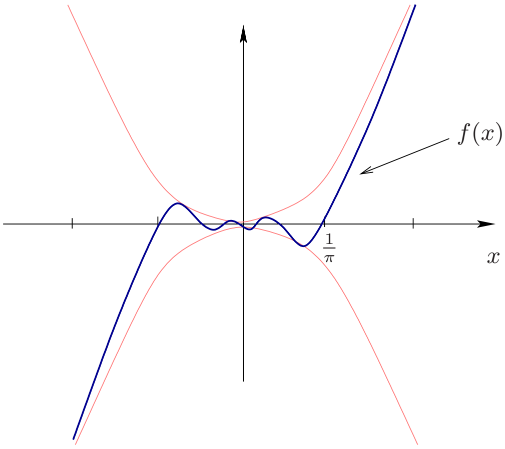


Fig. 2.1.1: A classic example.

The following theorem summarizes a basic property of Clarke gradient.

Theorem 2.1.2 [5] *Let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitzian near $x \in \mathbb{R}^n$ and let Ω be any set of Lebesgue measure zero. Then*

$$\partial f(x) = \text{co}\{\lim_{i\to\infty} \nabla f(x_i) : \quad \lim_{i\to\infty} x_i = x,\ x_i \notin \Omega,\ x_i \notin G_f\} \tag{2.1.2a}$$

$$f^o(x,d) = \limsup_{\substack{y\to x \\ y\notin \Omega\cup G_f}} \nabla f(y)^T d \qquad \forall\ d \in \mathbb{R}^n\ , \tag{2.1.2b}$$

where $G_f \overset{\triangle}{=} \{x \in \mathbb{R}^n :\quad f$ is not differentiable$\}$. A vector $g \in \partial f(x)$ is called a subgradient of f at x.

We remark (Theorem I.2.2.3) that for convex functions the Clarke gradient coincides with the subdifferential. Consequently each convex function is “regular”. Clarke’s theory provides a very powerful theoretical tool. On the other hand the practical use of concepts such as Clarke derivative in designing numerical optimization algorithms appears conditioned by the fact that the regularity assumption is very strong[4].

[4] “The Clarke derivative is too rough to be used for local approximations since it often gives unrealistic results” [6, p.101].

Example 2.1.2 *Consider the simple function, displayed in Fig. 2.1.2,*

$$f(x) = \begin{cases} \min\{x, 2-x\} & 0 \leq x \leq 2 \\ 0 & \textit{otherwise} \end{cases}$$

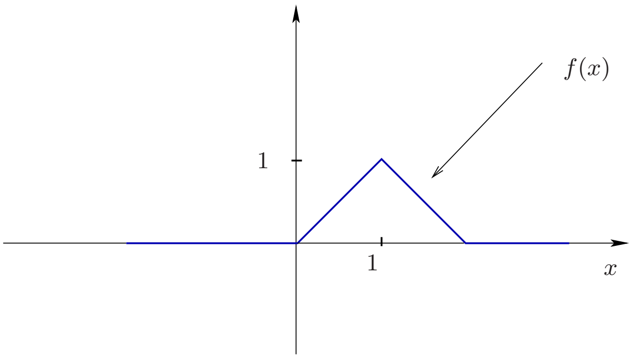


Fig. 2.1.2: A non regular function.

Let $\bar{x} = 1$. *By Theorem 2.1.2, we have* $\partial f(\bar{x}) = [-1, 1]$ *and*

$$f^o(\bar{x}, d) = |d| \quad \textit{and} \quad f'(x, d) = -|d| \qquad \forall\ d \in \mathbb{R}$$

Thus f *is not regular.*

Now we state a necessary condition of optimality for locally Lipschitzian functions used in the analysis of convergence of algorithms which tackle nonsmooth minimization problems.

Theorem 2.1.3 (Stationarity condition)[22] *Let* $f : \mathbb{R}^n \to \mathbb{R}$ *be Lipschitzian near* $x^* \in \mathbb{R}^n$. *If* f *attains its local minimum at* x^*, *then*

$$0 \in \partial f(x^*). \tag{2.1.3}$$

2.2 Goldstein ϵ-Subdifferential

Now we define the Goldstein ϵ-subdifferential.

Definition 2.2.1 (Goldstein ϵ-subdifferential)[22] *Let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitzian near $x \in \mathbb{R}^n$ and $\epsilon \geq 0$. Then the Goldstein ϵ-subdifferential of f at x, denoted $\partial_\epsilon^G f(x)$, is given by*

$$\mathrm{co}\{\partial f(y) : \quad \|y - x\| \leq \epsilon\} \ .$$

Each element $g \in \partial_\epsilon^G f(x)$ is called an ϵ-subgradient of f at x.

It is clear that $\partial f(x) \subset \partial_{\epsilon_1}^G f(x) \subset \partial_{\epsilon_2}^G f(x)$ for $0 \leq \epsilon_1 \leq \epsilon_2$. Of course if we substitute $\partial f(x)$ with $\partial_\epsilon^G f(x)$, $\epsilon \geq 0$, into the statement of Theorem 2.1.3 the result is valid again.

2.3 Demyanov and Rubinov Quasidifferential

Quasidifferential calculus was introduced in [6] by V.F. Demyanov and A.M. Rubinov. It provides an interesting alternative to Clarke gradient, based on the definition of two sets at any point x, namely $\underline{\partial} f(x)$ and $\overline{\partial} f(x)$, the subdifferential and the superdifferential respectively. The interested reader is referred to the complete treatment provided in [7].

3 Mean Value Theorems and Semismoothness

Now we deal with other theoretical tools used in the analysis of convergence of some numerical optimization algorithms.

3.1 Mean Value Theorems

Now we present two mean value theorems: one for directionally differentiable functions and the other for locally Lipschitzian functions.

Lemma 3.1.1 [7] *Let f be a real-valued function defined and continuous on an interval $[y, z]$ and suppose that at each point $x \in [y, z]$ there exists the right-hand-side derivative:*

$$f'_+(x) \triangleq \lim_{d \downarrow 0} \frac{f(x+d) - f(x)}{d} .$$

If

$$f'_+(x) \geq 0 \quad \forall \ x \in [y, z] \ , \tag{3.1.1}$$

then

$$f(z) \geq f(y) \ . \tag{3.1.2}$$

Proof. For any given $\epsilon > 0$, consider the set

$$\Omega \overset{\triangle}{=} \{x \in [y,z] : \quad f(\xi) - f(y) \geq -\epsilon(\xi - y) \quad \forall\ \xi \in [y,x]\} \ . \tag{3.1.3}$$

We prove that Ω coincides with $[y,z]$. It is clear that $y \in \Omega$; in fact

$$f(y) - f(y) = -\epsilon(y - y).$$

Let $x \in \Omega$; so we have $[y,x] \subset A$. Two cases can occur:

(i) $\Omega = [y, \gamma)$

(ii) $\Omega = [y, \gamma]$

for some $\gamma \in [y,z]$. From (3.1.3), we have

$$f(\xi) - f(y) \geq -\epsilon(\xi - y) \quad \forall\ \xi \in [y, \gamma)$$

and, since f is continuous on $[y,z]$,

$$f(\gamma) - f(y) = \lim_{\xi \downarrow \gamma} f(\xi) - f(y) \geq \lim_{\xi \downarrow \gamma} -\epsilon(\xi - y) = -\epsilon(\gamma - y) \ . \tag{3.1.4}$$

Consequently Ω coincides with $[y, \gamma]$ and the case (i) is impossible.

We prove, by contradiction, that $\gamma = z$, i.e. $\Omega = [y,z]$. Suppose in fact that $\gamma < z$, then for all $\delta > 0$ there exists an $\alpha_\delta \in (0, \delta]$ such that

$$f(\gamma + \alpha_\delta) - f(y) < -\epsilon(\gamma + \alpha_\delta - y) \ . \tag{3.1.5}$$

It follows from (3.1.4) and (3.1.5) that

$$\begin{array}{rcl} f(\gamma + \alpha_\delta) - f(\gamma) & = & f(\gamma + \alpha_\delta) - f(y) - (f(\gamma) - f(y)) \\ & < & -\epsilon(\gamma + \alpha_\delta - y) + \epsilon(\gamma - y) \\ & = & -\epsilon \alpha_\delta \end{array}$$

This means that

$$f'_+(\gamma) \leq -\epsilon,$$

which contradicts (3.1.1). Thus $z \in \Omega$ and therefore

$$f(z) - f(y) \geq -\epsilon(z - y)$$

which in turn, taking into account that ϵ is an arbitrary positive number, implies (3.1.2).

□

Corollary 3.1.2 [7] *Let f be a real-valued function defined and continuous on an interval $[y, z]$ and suppose that at each point $x \in [y, z]$ there exists right-hand-side derivative. Then*

$$m(z - y) \leq f(z) - f(y) \leq M(z - y),$$

where $m \triangleq \inf_{x \in [y,z]} h'_+(x)$ e $M \triangleq \sup_{x \in [y,z]} f'_+(x)$.

Proof. We apply Lemma 3.1.1 to functions

$$f_1(x) \triangleq Mx - f(x), \qquad f_2(x) \triangleq f(x) - mx \ .$$

□

Finally we show the mean value theorem for directionally differentiable functions.

Theorem 3.1.3 (Mean value theorem I)[7] *Let f be a real-valued function defined and continuous on an open set $A \subset \mathbb{R}^n$. Fix $x_0 \in A$, $t_0 \in \mathbb{R}$ and $d \in \mathbb{R}^n$ such that*

$$\mathcal{C} \triangleq [x_0, \ x_0 + t_0 d] \subset A \ .$$

Suppose that there exists the directional derivative of f at each point $x \in \mathcal{C}$ along the direction d. Let $m \triangleq \inf_{x \in \mathcal{C}} f'(x, d)$ and $M \triangleq \sup_{x \in \mathcal{C}} f'(x, d)$. Then

$$f(x_0 + t_0 d) = f(x_0) + ct_0,$$

for some $c \in [m, M]$.

Proof. Let $\phi(t) \triangleq f(x_0 + td)$ be defined on $[0, \ t_0]$. Then the thesis follows by applying Corollary 3.1.2.

□

We report the mean value theorem for locally Lipschitzian functions without the proof.

Theorem 3.1.4 (Mean value theorem II: Lebourg)[5] *Let x and y be points in $\mathbb{R}^n$, and suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is Lipschitzian on an open set containing the line $[x, y]$. Then there exists a vector $g \in \partial f(x + t(y - x))$ with $t \in (0, 1)$ such that*

$$f(y) - f(x) = g^T(y - x) \ .$$

3.2 Semismoothness and Weak Semismoothness

The notion of semismoothness was originally introduced by R. Mifflin [24]. Obviously, convex functions and smooth functions are semismooth.

Definition 3.2.1 (Semismoothness and weak semismoothness)[30] *Let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitzian near $x \in \mathbb{R}^n$. Suppose that*

$$\lim_{\substack{d' \to d \\ t \downarrow 0}} g(t)^T d' \quad \textit{(or } \lim_{t \downarrow 0} g(t)^T d\textit{)}$$

exists for all $d \in \mathbb{R}^n$, where $g(t) \in \partial f(x + td')$ (or $g(t) \in \partial f(x + td)$). Then f is said to be semismooth (or weakly semismooth) at x. Moreover f is called semismooth (or weakly semismooth) if it is semismooth (weakly semismooth) at each $x \in \mathbb{R}^n$.

Moreover if f is weakly semismooth, then it is locally Lipschitzian (see the proof of the Theorem I.1.2.2).

Proposition 3.2.1 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be weakly semismooth at $x \in \mathbb{R}^n$ and let d be any vector in $\mathbb{R}^n$. Then*

$$f'(x, d) = \lim_{t \downarrow 0} g(t)^T d$$

where $g(t) \in \partial f(x + td)$.

Proof. Let $\{\tau_k\}$ be a sequence such that $\lim_{k \to \infty} \tau_k = 0$. Then by virtue of Lebourg theorem

$$\lim_{\tau_k \downarrow 0} \frac{f(x + \tau_k d) - f(x)}{\tau_k} = \lim_{k \to \infty} g_k^T d, \tag{3.2.1}$$

where $g_k \in \partial f(x_k + t_k d)$ for some $t_k \in (0, \tau_k)$. Consequently the thesis follows.

□

4 Rademacher Theorem

Rademacher Theorem [33] (1919) is a theoretical tool largely used in nonsmooth optimization, since it explains that for a very large class of real-valued functions defined on $\mathbb{R}^n$ the "lacking in information" part of the space $\mathbb{R}^n$ (the nondifferentiable points) is a set of measure zero in the sense of Lebesgue measure.

4.1 Case 1): Functions of a real variable

The following theorem was proved first by H. Lebesgue (1904) for continuous monotonic functions and subsequently by F. Riesz (see [34]) for monotonic and not necessarily continuous functions.

Theorem 4.1.1 (Lebesgue's Theorem)[34] *Let f be a real-valued function defined and monotonic on an interval $[y, z]$. Then f possesses a (finite) derivative at every point $x \in [y, z]$ with the possible exception of the points x of a set of measure zero, or, as it is often phrased, almost everywhere.*

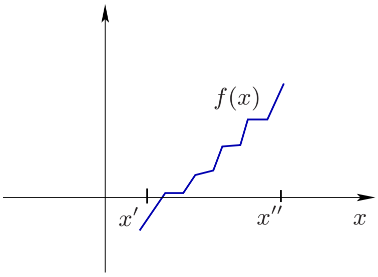


Fig. 4.1.1: Continuous monotonic functions

Remark 4.1.1 *The statement of Theorem 4.1.1 holds also for real-valued functions defined and monotonic on $\mathbb{R}$. By a simple trick, we prove this result. In partic-*

ular, we partition $\mathbb{R}$ into the intervals $[n, n+1)$, for $n \in \mathbb{Z}$. Then the derivative of f, by Theorem 4.1.1, exists everywhere on each interval $[n, n+1)$ and so, by applying (B.1b), it follows that f is differentiable on $\mathbb{R}$ a.e.

Definition 4.1.1 (Functions of bounded variation)[34] *Let f be a real-valued function defined on an interval $[y, z]$. The function f is said to be function of bounded variation, if the sum*

$$\sum_{i=1}^{n} |f(x_i) - f(x_{i-1})|$$

does not surpass a finite bound for any choice of the decomposition of $[y, z]$, denoted as $T(y, z) \overset{\triangle}{=} \{y = x_0 < x_1 < \ldots < x_n = z\}$.

The set of functions of bounded variation on $[y, z]$ is denoted by $\mathrm{BV}[y, z]$.

Theorem 4.1.2 [34] *Let $f \in \mathrm{BV}[y, z]$. Then f is the difference of two nondecreasing functions, i.e. there exist two nondecreasing functions f_1, f_2 such that $f(x) = f_1(x) - f_2(x)$ for every $x \in [y, z]$.*

Finally we show "Rademacher Theorem" for functions Lipschitzian and defined on an interval $[y, z]$ taking values in $\mathbb{R}$.

Proposition 4.1.3 *Let f be a real-valued function defined and Lipschitzian on an interval $[y, z]$. Then f is of bounded variation and almost everywhere differentiable on $[y, z]$.*

Proof. Indicate by L a Lipschitzian constant of f on $[y, z]$. By Definition C.1, we have

$$\sum_{i=1}^{n} |f(x_i) - f(x_{i-1})| \leq L \sum_{i=1}^{n} |x_i - x_{i-1}| = L(z - y).$$

It follows that $f \in \mathrm{BV}[y, z]$, which in turn, by both theorems 4.1.1 and 4.1.2, implies that f is differentiable on $[y, z]$ a.e.

□

Proposition 4.1.3 summarizes also Rademacher Theorem for real-valued Lipschitzian functions defined on the whole of real axis (see Remark 4.1.1).

Proposition 4.1.4 *Let f be a real-valued function defined and convex on an* *interval $[y,z]$. Then is differentiable at every point $x \in [y,z]$ except at most in a countable set of points.*

Proof. Let x be any point of (y,z). By considering that the slope-function

$$d \mapsto \frac{f(x+d)-f(x)}{d}$$

is increasing on $[y-x, z-x] \setminus \{0\}$ (see [15]), then the limits

$$f'_+(x) \triangleq \lim_{d \downarrow 0} \frac{f(x+d)-f(x)}{d} \quad \text{e} \quad f'_-(x) \triangleq \lim_{d \uparrow 0} \frac{f(x+d)-f(x)}{d}$$

exist.

From Axiom of Choice (see, e.g., [36]), it follows that $f'_+(x)$ and $f'_-(x)$ are continuous at each point of the domain except at most in a countable set of points. Taking into account that

$$f'_+(x) \leq f'_-(x+d) \leq f'_+(x+d), \quad d > 0,$$

the thesis follows by passing to the limit as $d \downarrow 0$.

□

4.2 Case 2): Functions of several real variables

The statement of Rademacher Theorem [33] is more general that well-known version used in nonsmooth optimization and here presented.

Lemma 4.2.1 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be Lipschitzian on $\mathbb{R}^n$ and let d be a given point in $\mathbb{R}^n$. Let τ be an arbitrary positive number. Then*[5] *the functions*

$$x \mapsto \overline{f}'(x,d), \quad x \mapsto \underline{f}'(x,d), \quad x \mapsto \overline{g}^x_\tau(d), \quad x \mapsto \underline{g}^x_\tau(d)$$

are measurable functions.

Theorem 4.2.2 (Rademacher Theorem)[25] *Let f be a real-valued function defined and Lipschitzian on a set $A \subset \mathbb{R}^n$. Then f is almost everywhere Fréchet differentiable on A.*

Proof.

[5] The functions considered in Lemma 4.2.1 are defined in (1.5.1)

To simplify the proof, we consider the case $A = \mathbb{R}^n$.

Consider the measure space $(\mathbb{R}^n, \mathcal{L}_n, \lambda_n)$. Let[6]

$$A_f \overset{\triangle}{=} \bigcup_{d \in \mathbb{R}^n} A_{f,d} \ ,$$

where[7] $A_{f,d} \overset{\triangle}{=} \{x \in \mathbb{R}^n : \ f^*(x,d) \text{ does not exists}\}$. Let C be a countable dense subset of $\mathbb{R}^n$ and let L be a Lipschitzian constant of f on $\mathbb{R}^n$.

Choose any $z \notin \bigcup\limits_{d \in C} A_{f,d}$. Let ϵ be an arbitrary positive number and let $\hat{d} \notin C$. From Definition A.7, we have that there exists $\bar{d} \in C$ such that

$$\|\hat{d} - \bar{d}\| \leq \frac{\epsilon}{L}$$

Taking into account that $d \mapsto \overline{f}'(x,d)$ and $d \mapsto \underline{f}'(x,d)$ are Lipschitzian functions (see [25]) with constant L on any compact set of $\mathbb{R}^n$, we have[8]

$$\left|\overline{f}'(z,\hat{d}) - f^*(z,\bar{d})\right| \leq L\|\hat{d} - \bar{d}\| \leq \epsilon$$
$$\left|\underline{f}'(z,\hat{d}) - f^*(z,\bar{d})\right| \leq L\|\hat{d} - \bar{d}\| \leq \epsilon$$

which in turn implies

$$-2\epsilon \leq \overline{f}'(z,\hat{d}) - \underline{f}'(z,\hat{d}) \leq 2\epsilon$$

Thus $\overline{f}'(z,\hat{d}) = \underline{f}'(z,\hat{d})$, and consequently $f^*(z,\hat{d})$ exists. In other words, we obtain

$$A_f = \bigcup_{d \in C} A_{f,d}.$$

Let $\phi(t) \overset{\triangle}{=} f(x_0 + td)$. By Proposition (4.1.3), we have that

$$A_{f,d} \cap \{x_0 + td : \ t \in \mathbb{R}\} = \{x_0 + td : \ \phi'(t) \text{ does not exists}\}$$

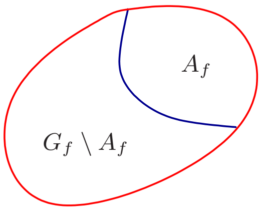
is a set of linear measure zero. By virtue of Fubini's Theorem (see, e.g., [34]), we have $\lambda_n(A_{f,d}) = 0$. Hence, taking into account that C is a countable set, we have

$$\lambda_n(A_f) = 0 \ \ .$$

[6] $f^*(x,d)$ is defined in (1.3.3)

[7] $A_{f,d}$ is a measurable set. See [14] for a discussion on how to prove this statement.

[8] The directional Dini derivatives exist everywhere.



$A_f = \{x \in \mathbb{R}^n : \ f^*(x,d)$ does not exists for any $d \in \mathbb{R}^n\}$

$G_f \setminus A_f = \{x \in \mathbb{R}^n : \ d \mapsto f^*(x,d)$ is not a linear function$\}$

Fig. 4.2.1: The entire space: $\mathbb{R}^n$.

Let $G_f \triangleq \{x \in \mathbb{R}^n : \ f$ is not Gâuteaux differentiable at $x\}$. In order to prove that f is almost everywhere Fréchet differentiable on $\mathbb{R}^n$, by Proposition 1.5.1 it is sufficient to show that f is almost everywhere Gâuteaux differentiable on $\mathbb{R}^n$ or equivalently $G_f \setminus A_f$ is a set of measure zero.

Choose any point $y \in G_f \setminus A_f$. Of course $d \mapsto f^*(y,d)$ is a homogeneous function, i.e. for all $d \in \mathbb{R}^n$

$$f^*(y,d) = tf^*(y,d) \quad \forall\, t \in \mathbb{R},$$

Hence the function $d \mapsto f^*(y,d)$ is not additive, i.e. there exist $d_1, d_2 \in \mathbb{R}^n$ satisfying the following relation

$$f^*(y,d_1) + f^*(y,d_2) - f^*(y,d_1+d_2) \neq 0 \ ,$$

and from the continuity of $d \mapsto f^*(y,d)$ there exist $v_1, v_2 \in C$ such that

$$f^*(y,v_1) + f^*(y,v_2) - f^*(y,v_1+v_2) \neq 0 \ . \tag{4.2.1}$$

Then consider the following measurable[9] sets

$$B(w_1,w_2,r_1,r_2) \triangleq \{x \notin A_f : \quad f^*(x,w_1) > r_1,\ f^*(x,w_2) > r_2,\ f^*(x,w_1+w_2) < r_1+r_2\}$$

$$B^*(w_1,w_2,r_1,r_2) \triangleq \{x \notin A_f : \quad f^*(x,w_1) < r_1,\ f^*(x,w_2) < r_2,\ f^*(x,w_1+w_2) > r_1+r_2\}$$

$$B(w_1,w_2,r_1,r_2,m) \triangleq \{x \in \mathbb{R}^n : \ \underline{g}^x_{\frac{1}{m}}(w_1) > r_1,\ \underline{g}^x_{\frac{1}{m}}(w_2) > r_2,\ \overline{g}^x_{\frac{1}{m}}(w_1+w_2) < r_1+r_2\}$$

Let us prove that $B(w_1,w_2,r_1,r_2)$ is a set of measure zero.

[9] See [14] for a discussion on how to prove this statement.

i) $w_1 + w_2 = 0$: $f^*(x, w_1) + f^*(x, w_1) = f^*(x, w_1 + w_2)$. From (B.1a), we have $\lambda_n \left(B(w_1, w_2, r_1, r_2)\right) = 0$.

ii) $w_1 + w_2 \neq 0$: Choose any $m \in \mathbb{N} \setminus \{0\}$. Let $T \triangleq B(w_1, w_2, r_1, r_2, m) \cap \{x_0 + t(w_1 + w_2) : \ t \in \mathbb{R}\}$. Let $x', x'' \in T$ be two points such that $x' \neq x''$. Moreover take $\bar{t} = \dfrac{\|x'' - x'\|}{\|w_1 + w_2\|}$.

Two cases can occur:

1. $x'' = x' + \bar{t}(w_1 + w_2)$;
2. $x' = x'' + \bar{t}(w_1 + w_2)$.

We can assume without loss of generality that the first case occurs. We show, by contradiction, that the following inequality holds:

$$\|x'' - x'\| \geq \frac{\|w_1 + w_2\|}{m} \ .$$

Suppose in fact that

$$\|x'' - x'\| < \frac{\|w_1 + w_2\|}{m}$$

holds. Thus, by the definition of the set $B(v_1, v_2, r_1, r_2, m)$, we have

$$f(x' + \bar{t}w_1) - f(x') > r_1 \bar{t} \tag{4.2.2a}$$

$$-f(x' + \bar{t}w_1) + f(x'') > r_2 \bar{t} \tag{4.2.2b}$$

$$f(x'') - f(x') < r_1 \bar{t} + r_2 \bar{t} \tag{4.2.2c}$$

which yields a contradiction. As consequence of Axiom of Choice (see, e.g., [36]), we have that T is countable, which in turn implies that T is a set of linear measure zero. From Fubini's Theorem, we have $\lambda_n(B(w_1, w_2, r_1, r_2, m)) = 0$.

Taking into account that

$$B(w_1, w_2, r_1, r_2) \subset \bigcup_{m=1}^{\infty} B(v_1, v_2, r_1, r_2, m) \ ,$$

we have $\lambda_n(B(w_1, w_2, r_1, r_2)) = 0$.

From corollary of Axiom of Archimedes (see, e.g., [36]), we have

$$G_f \setminus A_f = \bigcup_{\substack{r_1, r_2 \in \mathbb{Q} \\ w_1, w_2 \in C}} B(w_1, w_2, r_1, r_2) \cup B^*(w_1, w_2, r_1, r_2). \tag{4.2.3}$$

Since $B(w_1, w_2, r_1, r_2) = B^*(-w_1, -w_2, -r_1, -r_2)$, it follows that that $G_f \setminus A_f$ is a set of measure zero. Consequently

$$\mu(G_f) = \mu(G_f \setminus A_f) + \mu(A_f) = 0.$$

It is clear that the scheme of the proof is unchanged, if we remove the assumption $A = \mathbb{R}^n$.

□

The statement of Theorem 4.2.2 holds also for locally Lipschitzian functions (see Remark 4.1.1).

Chapter V

Algorithms for Nonsmooth Nonconvex Optimization

Introduction. In this chapter we present some numerical algorithms for nonsmooth nonconvex optimization. We describe first the bundle method BTNC [37] by H. Schramm and J. Zowe. This method consists essentially in adapting the bundle method BTC [37] for nonsmooth convex optimization to solve nonsmooth nonconvex problems. Then we describe two recent bundle methods, NCVX and DC-NCVX [10, 11] by A. Fuduli, M. Gaudioso and G. Giallombardo, designed for nonsmooth nonconvex optimization. Finally we describe the Gradient Sampling algorithm introduced in [3] by J. Burke, A. Lewis and L. Overton.

1 Some Bundle Methods

We consider the unconstrained minimization problem:

$$(P) \begin{cases} \min f(x) \\ \quad x \in \mathbb{R}^n \ , \end{cases} \tag{1.0.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is not necessarily differentiable.

We require that f is locally Lipschitzian, thus it is (Theorem IV.4.2.2) differentiable almost everywhere. We assume the existence of an oracle which, given any point $x \in \mathbb{R}^n$, calculates both the objective function value $f(x)$ and a subgradient $g \in \partial f(x)$, i.e. an element of the Clarke gradient. We assume also that a starting point, say x_1, is available and we indicate by y_k the stability center during the execution of k^{th} "main iteration".

1.1 BTC and BTNC Algorithms

We present both convex (BTC) and nonconvex (BTNC) versions of the approach by H. Schramm and J. Zowe. We describe first the convex version.

BTC Algorithm[37] We assume that f is convex. We remark (Theorem I.1.2.2) that if f is a convex function, then it is also locally Lipschitzian. This method is quite like Algorithm II.2.1.2, therefore we describe only the different steps, that is the k^{th} "main iteration" (steps 1 and 2 of the Algorithm II.2.1.2).

We set the following parameters:

- the stopping parameter ϵ and the descent parameter $m_1 \in (0,1)$;
- the safeguard parameters $\rho_{\min}$ and $\rho_{\max}$, $0 < \rho_{min} < \rho_{max}$;
- $\nu > 0$, $m_3 \in (0,1)$ and $m_2 \in (m_1, 1)$.

Algorithm 1.1.1 (k^{th} BTC *main iteration*)[37]

Step 1. *Set* $\rho^1 := \rho_{k-1}$. *Fix* $l^1 := \rho_{\min}$, $u^1 := \rho_{\max}$ *and* $i := 1$.

a) *Solve*[1] *either* (QP_{ρ^i}) *or* (DP_{ρ^i}) *and compute* $(d(\rho^i), v(\rho^i))$ *and* $\lambda(\rho^i)$. *If* $\rho^i \|d(\rho^i)\| \leq \epsilon$ *and* $-v(\rho^i) - \rho^i \|d(\rho^i)\|^2 \leq \epsilon$, *then stop:* y_k *is* ϵ*-optimal. Else put* $x^i := y_k + d(\rho^i)$, *evaluate* f *at* x^i *and compute* $g^i \in \partial f(x^i)$.

Step 2. *Two cases can occur:*

a) $f(x^i) < f(y_k) + m_1 v(\rho^i)$. *If* $g^{iT} d(\rho^i) \geq m_2 v(\rho^i)$ *or* $\rho^i \leq (\rho_{\min} + \nu)$, *then calculate*[2] g_p^k, α_p^k *and put* $x_{k+1} := x^i$, $g_{k+1} := g^i$. *Else put* $l^{i+1} := l^i$, $u^{i+1} := \rho^i$, *set* $\rho^{i+1} := 0.5(l^{i+1} + u^{i+1})$, *replace* i *by* $i+1$ *and go to step* 1*a*).

b) $f(x^i) \geq f(y_k) + m_1 v(\rho^i)$. *If*[3] $\alpha^{ik} \leq m_3 \alpha_p^{k-1}$ *or* $|f(y_k) - f(x^i)| \leq \|g_p^{k-1}\| + \alpha_p^{k-1}$ *or* $\rho^i \geq (\rho_{\max} - \nu)$, *then calculate* g_p^k, α_p^k, *put* $y_{k+1} := y_k$, $x_{k+1} := x^i$, $g_{k+1} := g^i$ *and*

[1] (QP_{ρ^i}) and (DP_{ρ^i}) are defined in §II.2.1

[2] g_p^k and α_p^k are defined in §II.2.1

[3] α^{ik} is $f(y_k) - [f(x^i) + g^{iT}(y_k - x^i)]$

go to step 4. Else put $l^{i+1} := \rho^i$, $u^{i+1} := u^i$, set $\rho^{i+1} := 0.5(l^{i+1} + u^{i+1})$, replace i by $i+1$ and go to step 1a).

We remark that if the algorithm enters step 2a), then the descent condition

$$f(x^i) < f(y_k) + m_1 v(\rho^i)$$

is fulfilled; whereas if the algorithm enters step 2b), then a cut for the subproblem (QP_{ρ^i}) is generated and thus the optimal solutions of $(QP_{\rho^{i+1}})$ and (QP_{ρ^i}) are not the same. The Algorithm 1.1.1 implements a quite standard strategy to adjust the proximity parameter ρ.

Theorem 1.1.2 [37] *Let $\epsilon = 0$ and let f be convex. Then the sequence $\{y_k\}$, generated by the BTC algorithm, converges to the infimum of f on $\mathbb{R}^n$ as $k \to \infty$.*

Remark 1.1.1 *For convex function f, since the subgradient inequality holds, we have*

$$\check{f}_k(y_k + d)\big|_{d=0} = \max_{j \in I_k} \{f(y_k) + g_j^T d - \alpha_j^k\}\Big|_{d=0} = f(y_k) + \max_{j \in I_k}\{-\alpha_j^k\} = f(y_k)$$

Hence it is guaranteed that function $\check{f}_k$ interpolates f at y_k.

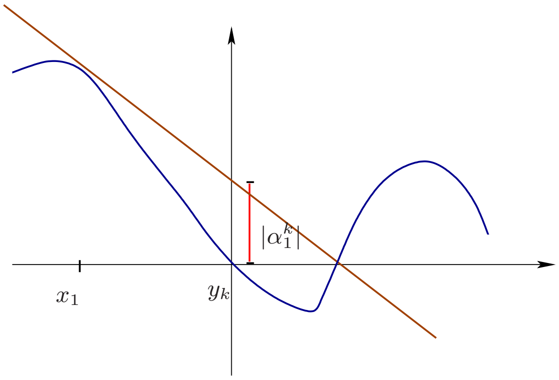


Fig. 1.1.1: $\alpha_1^k < 0$.

BTNC Algorithm[37] For nonconvex function f, the linearization error α_j^k, $j \in I_k$, could be negative, since the first order expansion at any point does not necessarily support from below the epigraph of the function, as shown in Fig. 1.1.1. Consequently the values of f and $\check{f}_k$ at y_k could be different. Moreover the cutting-plane function $\check{f}$ is not necessarily a lower approximation of the objective function f. Schramm and Zowe, in [37], adopt a technique also used by Lemaréchal in [18] in order to guarantee interpolation at y_k. In particular α_j^k, for all $j \in I_k$, is replaced by

$$\beta_j^k \overset{\triangle}{=} \max\{\alpha_j^k,\ c_0\|x_j - y_k\|^2\} \geq 0,$$

where c_0 is a fixed small positive number. Thus the modified cutting-plane approximation of f, at iteration k, is

$$\bar{f}_k(y_k + d) \overset{\triangle}{=} \max_{j \in I_k}\{f(y_k) + g_j^T d - \beta_j^k\}$$

Consequently $\bar{f}_k(y_k + d)\big|_{d=0} = f(y_k) + \max\limits_{j\ \in I_k}\{-\beta_j^k\} = f(y_k)$. Moreover both the subproblems (QP_{ρ^i}) and (DP_{ρ^i}) remain formally unchanged.

For nonconvex case, we assume that f is weakly semismooth as well, so that the problem of finding the scalar t at step 2bb) of Algorithm 1.1.3 is well-posed. The "main iterations" for convex and nonconvex cases are very much alike.

We set the following parameters:

- the stopping parameter ϵ and the descent parameter $m_1 \in (0, 1)$;
- $m_3 \in (0, 1),\ m_2 \in (m_1, 1)$ and $c_0 > 0$;

Algorithm 1.1.3 (k^{th} BTNC *main iteration*)
Step 1. *Set* $\rho^1 = \rho_{k-1}$. *Fix* $l^1 := 0,\ u^1 := \rho_{max}$ *and* $i := 1$.

1a) *Solve either* (QP_{ρ^i}) *or* (DP_{ρ^i}) *and compute* $(d(\rho^i), v(\rho^i))$ *and* $\lambda(\rho^i)$. *If* $\rho^i\|d(\rho^i)\| \leq \epsilon$ *and* $-v(\rho^i) - \rho^i\|d(\rho^i)\|^2 \leq \epsilon$, *then stop*[4]. *Else put* $x^i := y_k + d(\rho^i)$, *evaluate* f *at* x^i *and compute* $g^i \in \partial f(x^i)$.

Step 2. *Two cases can occur:*

[4] $\|g_p^k\| \leq \epsilon, \quad g_p^k \in \text{co}\left\{\partial_\epsilon f(x_j) : \quad \|x_j - y_k\| \leq \sqrt{\frac{\epsilon}{c_0\lambda_j}}\right\}$

2a) $f(x^i) < f(y_k) + m_1 v(\rho^i)$. *Calculate*[5] $g_p^k,\ \beta_p^k$ *and put* $x_{k+1} := x^i,\ g_{k+1} := g^i$.

2b) $f(x^i) \geq f(y_k) + m_1 v(\rho^i)$. *Four cases can occur*[6]*:*

2ba) $\left[\beta^{ik} \leq m_3 \beta_p^{k-1}\right.$ *or* $\left.|f(y_k) - f(x^i)| \leq \|g_p^{k-1}\| + \beta_p^{k-1}\right]$ *and* $g^{iT} d(\rho^i) - \beta^{ik} \geq m_2 v(\rho^i)$. *Calculate* $g_p^k,\ \beta_p^k$*, put* $y_{k+1} := y_k,\ x_{k+1} := x^i,\ g_{k+1} := g^i$ *and go to step* 4.

2bb) $g^{iT} d(\rho^i) - \beta^{ik} < m_2 v(\rho^i)$ *and* $|f(y_k) - f(x^i)| \leq \|g_p^{k-1}\| + \beta_p^{k-1}$. *Calculate* $g_p^k,\ \beta_p^k$ *and find a scalar* $t \in (0,1)$ *such that either* $x_{k+1} := x^i + td(\rho^i)$ *satisfies a descent condition (serious step: updating stability center) or a cut for* (QP_{ρ^i}) *is generated (null step:* $y_{k+1} := y_k$ *and go to step* 4*).*

2bc) $g^{iT} d(\rho^i) - \beta^{ik} < m_2 v(\rho^i)$ *and* $\beta^{ik} \leq m_3 \beta_p^{k-1}$. *Put* $l^{i+1} := \rho^i,\ u^{i+1} := u^i$*, set* $\rho^{i+1} := 0.5(l^{i+1} + u^{i+1})$*, replace* i *by* $i+1$ *and go to step* 1*a).*

2bd) $\beta^{ik} > m_3 \beta_p^{k-1}$ *and* $|f(y_k) - f(x^i)| > \|g_p^{k-1}\| + \beta_p^{k-1}$. *Put* $l^{i+1} := \rho^i,\ u^{i+1} := u^i$*, set* $\rho^{i+1} := 0.5(l^{i+1} + u^{i+1})$*, replace* i *by* $i+1$ *and go to step* 1*a).*

We observe that at step 2ba) the condition

$$g^{iT} d(\rho^i) - \beta^{ik} \geq m_2 v(\rho^i)$$

assures that the solution of (QP_{ρ^i}), generated at the previous iteration, is cut away.

The algorithm BTNC is an adaptation of the convex bundle method BTC for dealing with nonconvex functions. The definition of the local model of the objective function f of (P) is somehow arbitrary, due to the possible "upward" translation of some of the affine pieces.

Theorem 1.1.4 [37] *Let* $\epsilon = 0$ *and let* f *be weakly semismooth and bounded below. If the sequence* $\{y_k\}$ *generated by the BTNC algorithm is bounded, then it converges to* y_k^* *such that* $0 \in \partial f(y_k^*)$.

We remark that the stopping condition at step 1a) is a very rough approximate stationarity condition.

1.2 NCVX and DC-NCVX Algorithms

We present two numerical algorithms by A. Fuduli, M. Gaudioso and G. Giallombardo designed for nonconvex case and reported in [10, 11].

[5] $\beta_p^k \overset{\triangle}{=} \lambda(\rho)^T \beta^k$

[6] β^{ik} is equal to $\max\{f(y_k) - [f(x^i) + g^{iT}(y_k - x^i)],\ c_0\|x^i - y_k\|\}$

We assume that the sublevel set

$$\mathcal{F}_1 \stackrel{\triangle}{=} \{x :\in \mathcal{R}^n : \quad f(x) \leq f(x_1)\}$$

is compact. As usual we denote by y_k the stability center during the execution of the k^{th} "main iteration".

NCVX Algorithm[10] The originality of this approach is contained in the bundle splitting technique, i.e. the index set I_k is divided into two sets I_k^+ and I_k^- on the basis of the sign of the linearization error. If the objective function is nonconvex, it is essential the way to handle the linearization error α_j^k, $j \in I_k$. In particular, I_k^+ and I_k^- are defined as

$$I_k^+ \stackrel{\triangle}{=} \{j : \quad \alpha_j^k \geq 0\} \quad \text{and} \quad I_k^- \stackrel{\triangle}{=} \{j : \quad \alpha_j^k < 0\} \ , \tag{1.2.1}$$

The bundles corresponding to the index sets I_k^+ and I_k^- are characterized by points that exhibit, respectively, a "convex behavior" and a "concave behavior" relative to y_k. We observe that I_k^+ is never empty, as at least the index corresponding to the stability center belongs to the I_k^+.

Let $h(d) \stackrel{\triangle}{=} f(y_k + d) - f(y_k)$ be the difference function. Then we consider two polyhedral approximations of h:

$$\Delta^+(d) \stackrel{\triangle}{=} \max_{j \ \in I_k^+} \left\{g_j^T d - \alpha_j^k\right\}$$

and

$$\Delta^-(d) \stackrel{\triangle}{=} \min_{j \ \in I_k^-} \left\{g_j^T d - \alpha_j^k\right\} .$$

We observe that $h(0) = \Delta^+(0)$ and, when I_k^- is nonempty, $\Delta^+(0) < \Delta^-(0)$. Consequently

$$\mathcal{C} \stackrel{\triangle}{=} \{d : \ \Delta^+(d) \leq \Delta^-(d)\}$$

denotes a kind of trust region model. The core of NCVX algorithm consists, as shown in Fig. 1.2.1, in finding a tentative stepsize by solving the following problem:

$$d(\rho) = \operatorname*{argmin}_{d \in \mathcal{C}} \Delta^+(d) + \frac{1}{2}\rho\|d\|^2 \ , \tag{1.2.2}$$

which, by introducing the scalar variable v, can be rewritten as a quadratic program

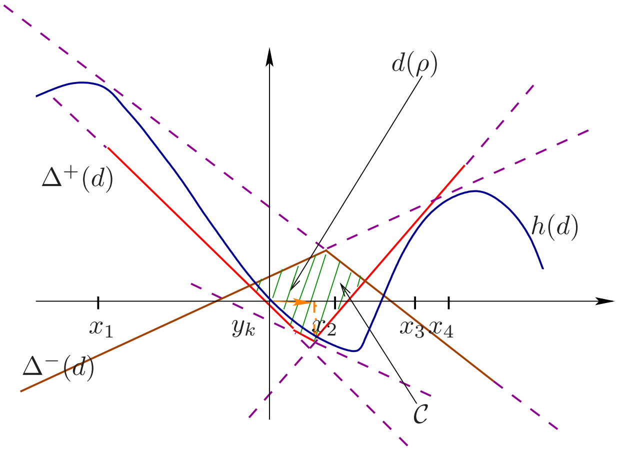


Fig. 1.2.1: NCVX approach.

of the form:

$$(QP_\rho)\begin{cases}\min\limits_{v,d}\ v+\dfrac{1}{2}\rho\|d\|^2\\ \quad v\geq g_j^Td-\alpha_j^k \quad \forall\ j\in I_k^+\\ \quad v\leq g_j^Td-\alpha_j^k \quad \forall\ j\in I_k^-\ .\end{cases} \tag{1.2.3}$$

The dual of $(QP\rho)$ can be written in the form:

$$(DP_\rho)\begin{cases}\min\limits_{\lambda\geq 0,\mu\geq 0} \quad \dfrac{1}{2\rho}\|G_+\lambda-G_-\mu\|^2+\lambda^T\alpha_+^k-\mu^T\alpha_-^k\\ \\ \qquad e^T\lambda-e^T\mu=1\ ,\end{cases}$$

where G_+ and G_- are matrices whose columns are, respectively, the vectors g_j, $j\in I_k^+$, and g_j, $j\in I_k^-$. Analogously, the terms α_j^k, $j\in I_k^+$, and α_j^k, $j\in I_k^-$, are grouped into the vectors α_+^k and α_-^k, respectively.

The optimal primal solution $(d(\rho),v(\rho))$ is related to the optimal dual solution $(\lambda(\rho),\mu(\rho))$ by the following formulae:

$$d(\rho)=\frac{1}{\rho}\left(-G_+\lambda(\rho)+G_-\mu(\rho)\right) \tag{1.2.4a}$$

$$v(\rho)=-\frac{1}{\rho}\left\|G_+\lambda(\rho)-G_-\mu(\rho)\right\|^2-\lambda(\rho)^T\alpha_+^k+\mu(\rho)^T\alpha_-^k\ . \tag{1.2.4b}$$

For nonconvex case, a small α_j^k, $j \in I_k^+ \cup I_k^-$, does not imply that the corresponding point x_j is near the stability center y_k. Thus the bundle is enriched by

$$a_j^k \triangleq \|x_j - y_k\|, \quad \forall\, j \in I_k^+ \cup I_k^-.$$

The bundle of available information used in [10, 11] is the set

$$\left\{(x_j, f(x_j), g_j, \alpha_j^k, a_j^k) : \quad j \in I_k^+ \cup I_k^- \right\}.$$

Now we describe the NCVX algorithm based on repeatedly solving problem (QP_ρ), or equivalently (DP_ρ). The kernel of the algorithm is the "main iteration", i.e. the set of steps where the stability center remains unchanged.

The whole algorithm can be summarized as follows:

Algorithm 1.2.1 (NCVX: algorithm outline)

1. *Initialization.*
2. *Execute the k^{th} "main iteration".*
3. *Bundle updating. Replace k by $k+1$ and return to 2.*

We remark that the updating of the bundle is necessary since the quantities α_j^k and a_j^k, $j \in I_k^+ \cup I_k^-$, are dependent on the stability center y_k.

The initialization of the algorithm requires a starting point $x_1 \in \mathbb{R}^n$. The initial stability center is set equal to x_1. The initial bundle is made up by just one element $(x_1, f(x_1), g_1, 0, 0)$, where $g_1 \in \partial f(x_1)$. Consequently I_1^- is the empty set, while I_1^+ is a singleton. The following global parameters are to be set:

- the stationarity tolerance $\eta > 0$ and the distance parameter $\epsilon > 0$;
- the descent parameter $m_1 \in (0,1)$ and the cut parameter $m_2 \in (m_1, 1)$;
- the reduction parameter $r \in (0,1)$, and the increase parameter $R > 1$.

The following local parameters are set each time the "main iteration" is entered:

- the proximity measure $\theta > 0$;
- the safeguard parameters $\rho_{\min}$ and $\rho_{\max}$, $0 < \rho_{min} < \rho_{max}$.

Two exits from the "main iteration" may occur:

(i) termination of the whole algorithm, due to satisfaction of an approximate stationarity condition;

(ii) update of the stability center, due to satisfaction of a sufficient decrease condition.

Algorithm 1.2.2 (k^{th} NCVX *main iteration*)

Step 1. *If* $\|g(y_k)\| \leq \eta$ *then stop.*

Set

$$\rho_{\max} := \frac{2R}{\epsilon}\|g(y_k)\|, \quad \rho_{\min} := r\rho_{\max}, \quad \theta := \frac{\rho_{\min}\eta}{R} \ .$$

Step 2. *Choose* $\hat{\rho}$ *equal to the maximum value of* $\rho \in [\rho_{\min}, \rho_{\max}]$ *such that:*

$$f(y_k + d(\rho)) > f(y_k) + m_1 v(\rho)$$

if such ρ *does exist. Otherwise set* $\hat{\rho} := \rho_{\min}$. *If* $\|d(\hat{\rho})\| > \theta$ *go to* 4.

Step 3. *Set*

$$I_k^+ := I_k^+ \setminus \{j \in I_k^+ : \quad a_j^k > \epsilon\}$$

and

$$I_k^- := I_k^- \setminus \{j \in I_k^- : \quad a_j^k > \epsilon\} \ .$$

Calculate

$$\|g^*\| = \min_{g \in \text{co}\{g_j:\ j \in I_k^+\}} \|g\| \ .$$

If $\|g^*\| \leq \eta$ *then stop. Else set* $\rho_{\min} := \rho_{\min} + r(\rho_{\max} - \rho_{\min})$ *and go to step* 2.

Step 4. *Set* $\hat{x} := y_k + d(\hat{\rho})$, *calculate* $\hat{g} \in \partial f(\hat{x})$ *and set*

$$\hat{\alpha} := f(y_k) - f(\hat{x}) + \hat{g}^T d(\hat{\rho}).$$

Step 5. a) *If* $\hat{\alpha} < 0$ *and* $\|d(\hat{\rho})\| > \epsilon$, *then insert the element* $(\hat{x}, f(\hat{x}), \hat{g}, \hat{\alpha}, \|d(\hat{\rho})\|)$ *into the bundle for an appropriate value of* $j \in I_k^-$ *and set* $\hat{\rho} := \hat{\rho} + r(\rho_{\max} - \hat{\rho})$.

b) *Else, if* $\hat{g}^T d(\hat{\rho}) \geq m_2 v(\hat{\rho})$ *then insert the element* $(\hat{x}, f(\hat{x}), \hat{g}, \max(0, \hat{\alpha}), \|d(\hat{\rho})\|)$ *into the bundle for an appropriate value of* $j \in I_k^+$.

c) *Else find a scalar* $t \in (0,1)$ *such that* $g(t) \in \partial f(y_k + td(\hat{\rho}))$ *satisfies the condition* $g(t)^T d(\hat{\rho}) \geq m_2 v(\hat{\rho})$ *and insert the element* $(y_k + td(\hat{\rho}), f(y_k + td(\hat{\rho})), g(t), \max(0, \alpha_t),$

$t\|d(\hat\rho)\|)$ *into the bundle for an appropriate value of* $j \in I_k^+$, *where* $\alpha_t = f(y_k) - f(y_k + td(\hat\rho)) + tg(t)^T d(\hat\rho)$.

Step 6. *If* $\|d(\hat\rho)\| \leq \theta$ *go to step* 3. *If*

$$f(\hat x) \leq f(y_k) + m_1 v(\hat\rho) \ , \tag{1.2.5}$$

set the new stability center $y_{k+1} := \hat x$ *and exit from the "main iteration".*

Step 7. *Solve* $QP(\hat\rho)$, *or equivalently* $DP(\hat\rho)$, *obtain both the primal and the dual optimal solution* $(v(\hat\rho), d(\hat\rho))$ *and* $(\lambda(\hat\rho), \mu(\hat\rho))$, *and go to step* 4.

We assume that f is weakly semismooth, so that the problem of finding the scalar t at step 5c) of Algorithm 1.2.2 is well-posed.

The separate handling of the points of bundle, according to sign of linearization error, is one of the greatest strengths of this approach. Consequently, this algorithm is not a simple adaptation of a convex bundle algorithm; it takes into account the nonconvex nature of the objective function f of (P).

The main disadvantage of this approach is that in the objective function of (1.2.2) only the lower polyhedral approximation Δ^+ is present, even though the upper polyhedral approximation is useful to define the feasible set of (1.2.2). Furthermore there is the danger of solving several subproblems (QP_ρ) at step 2.

Theorem 1.2.3 [10] *For any* $\epsilon > 0$ *and* $\eta > 0$, *NCVX algorithm stops in a finite number of "main iterations" at a stability center* y_k^* *satisfying the approximate stationarity condition*

$$\|g^*\| \leq \eta \qquad \text{with } g^* \in \partial_\epsilon^G f(y_k^*) \ . \tag{1.2.6}$$

DC-NCVX Algorithm[11] Now we utilize the technique of handling of the bundle used in NCVX, but we define a new local model of the objective function f of (P). In particular, the index sets I_k^+ and I_k^- are defined as follows

$$I_k^+ \triangleq \{j : \quad \alpha_j^k \geq 0\} \quad \text{and} \quad I_k^- \triangleq \{j : \quad \alpha_j^k \leq 0\} \ . \tag{1.2.7}$$

We remark that the sets I_k^+ and I_k^- are not disjoint; in fact at least the index corresponding to the stability center y_k belongs to both I_k^+ and I_k^-. Consequently $\Delta^+(d) \geq \Delta^-(d)$ for all $d \in \mathbb{R}^n$.

Let $\Delta_p(d) \overset{\triangle}{=} p\Delta^+(d) + (1-p)\Delta^-(d)$, for some $p \in (0,1)$.

The core of DC-NCVX algorithm consists, as displayed in Fig. 1.2.2, in finding a tentative stepsize by solving the following problem:

$$d_p(\rho) = \operatorname*{argmin}_{d\in\mathbb{R}^n} \Delta_p(d) + \frac{1}{2}\rho\|d\|^2 \tag{1.2.8}$$

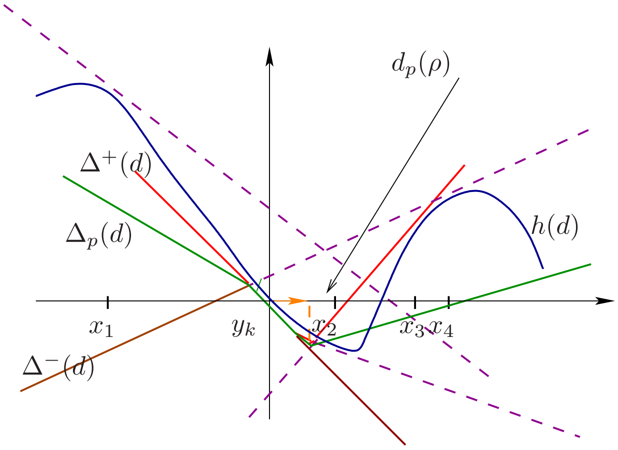


Fig. 1.2.2: DC-NCVX approach.

See [11] for a discussion on how to find a global optimal solution of the previous problem.

In this approach a relevant role is played also by the strictly convex program

$$d(\rho) = \operatorname*{argmin}_{d\in\mathbb{R}^n} \Delta^+(d) + \frac{1}{2}\rho\|d\|^2\ , \tag{1.2.9}$$

which is equivalent to the following quadratic programming problem

$$(QP_\rho)\begin{cases} \min\limits_{v,d} \quad v + \dfrac{1}{2}\rho\|d\|^2 \\ \\ \qquad v \geq g_j^T d - \alpha_j^k \quad j \in I_k^+\ . \end{cases}$$

The dual of $QP(\rho)$ is

$$(DP_\rho)\begin{cases} \min\limits_{\lambda\geq 0} & \dfrac{1}{2\rho}\|G_+\lambda\|^2+\lambda^T\alpha_+^k \\ & \\ & e^T\lambda = 1\ , \end{cases}$$

where G_+ and α_+^k are the quantities previously defined.

The optimal primal solution $(d(\rho), v(\rho))$ is related to the optimal dual solution $\lambda(\rho)$ by the following formulae:

$$d(\rho) = -\frac{1}{\rho}G_+\lambda(\rho) \tag{1.2.10a}$$

$$v(\rho) = -\frac{1}{\rho}\|G_+\lambda(\rho)\|^2 - \lambda(\rho)^T\alpha_+^k\ , \tag{1.2.10b}$$

where $v(\rho) = \Delta^+(d(\rho))$.

The initialization of the algorithm requires a starting point $x_1 \in \mathbb{R}^n$. The initial stability center is set equal to x_1. The initial bundle is made up by just one element $(x_1, f(x_1), g_1, 0, 0)$, where $g_1 \in \partial f(x_1)$. The corresponding index is put in both I_1^+ and I_1^-, which are consequently both a singleton. The scheme of whole algorithm coincides with Algorithm 1.2.1.

The following global parameters are to be set:

- the stationarity tolerance $\eta > 0$ and the distance parameter $\epsilon > 0$;
- the descent parameter $m \in (0, 1)$;
- the reduction parameter $r \in (0, 1)$ and the increase parameter $R > 1$;
- the initial balance parameter $p_0 \in (0, 1)$.

The following local parameters are set each time the "main iteration" is entered:

- the descent threshold parameter $\xi > 0$
- the safeguard parameters $\rho_{\min}$ and $\rho_{\max}$, $0 < \rho_{min} < \rho_{max}$
- the linearization error threshold parameter $\sigma > 0$ and the approximation parameter $\gamma > 0$;

- the balance parameter $p := p_0$.

The following conditions on the parameters are imposed during the "main iteration":

$$\rho_{\max} > \frac{2}{\epsilon}\|g_{I_k^-}\|, \tag{1.2.11}$$

where $\|g_{I_k^-}\| \overset{\triangle}{=} \max\limits_{j\ \in I_k^-}\{\|g_j\|\}$ and

$$\eta \geq \sqrt{\rho_{\max}(\xi+\gamma)}\ . \tag{1.2.12}$$

We assume that f is weakly semismooth, so that the problem of finding the scalar t at step 5d) of Algorithm 1.2.4 is well-posed.

Algorithm 1.2.4 (k^{th} DC-NCVX *main iteration*)
Step 1. *If* $\|g(y_k)\| \leq \eta$ *then stop.*
Step 2. *Set* $\rho_{\max} := \max\left\{\rho_{\max},\ \frac{2R}{\epsilon}\|g_{I_k^-}\|\right\}$, $\xi := \frac{\eta^2}{2\rho_{\max}}$, $\gamma := \xi$ *and* $\sigma := \gamma$. *Select* $\rho \in (\rho_{min}, \rho_{max})$.
Step 3. *Calculate* $d_p(\rho)$ *and* $d(\rho)$ *by solving, respectively, (1.2.8) and (1.2.9).*

If $\Delta^+(d_p(\rho)) \leq -\xi$ *go to* 4. *If* $\Delta^+(d_p(\rho)) - \Delta^+(d(\rho)) > \gamma$, *then set* $p := p + r(1-p)$ *and return to step* 2. *Else go to step* 7 .

Step 4. *Set* $\hat{x} := y_k + d_p(\rho)$. *If*

$$f(\hat{x}) \leq f(y_k) + m\Delta^+(d_p(\rho))$$

set the new stability center $y_k := \hat{x}$ *and exit from the main iteration.*

Step 5. *Calculate* $\hat{g} \in \partial f(\hat{x})$ *and set*

$$\hat{\alpha} := f(y_k) - f(\hat{x}) + \hat{g}^T d_p(\rho)\ .$$

Four cases can occur:

5a) $\hat{\alpha} \leq -\sigma$ *and* $\|d_p(\rho)\| > \epsilon$. *Set* $\rho := \rho + r(\rho_{\max} - \rho)$ *and return to step* 3.

5b) $\hat{\alpha} \geq \sigma$. *Insert the element* $(\hat{x}, f(\hat{x}), \hat{g}, \hat{\alpha}, \|d_p(\rho)\|)$ *into the bundle for an appropriate value of* $j \in I_k^+$ *and return to step* 3.

5c) $0 \leq \hat{\alpha} < \sigma$. *Insert the element* $(\hat{x}, f(\hat{x}), \hat{g}, 0, \|d_p(\rho)\|)$ *into the bundle twice, for two appropriate values of the indices one belonging to* I_k^+ *and the other to* I_k^-.

5d) *($\hat{\alpha} \leq -\sigma$ and $\|d_p(\rho)\| \leq \epsilon$) or ($-\sigma < \hat{\alpha} < 0$). Find a scalar $t \in (0,1)$ such that $g(t) \in \partial f(y_k + td_p(\rho))$ satisfies the condition*

$$g(t)^T d_p(\rho) \geq m\Delta^+(d_p(\rho)) \ , \tag{1.2.13}$$

insert the element $(y_k + td_p(\rho), f(y_k + td_p(\rho)), g(t), 0, t\|d_p(\rho)\|)$ into the bundle twice, for two appropriate values of the indices one belonging to I_k^+ and the other to I_k^-.

Step 6. *Set $\rho_{\max} := \max\left\{\rho_{\max},\ \frac{2R}{\epsilon}\|g_{I_k^-}\|\right\}$, $\xi := \frac{\eta^2}{2\rho_{\max}}$, $\gamma := \xi$, $\sigma := \gamma$ and return to step* 3.

Step 7. *Set*

$$I_k^+ := I_k^+ \setminus \{j \in I_k^+ : \quad a_j^k > \epsilon\}$$

and

$$I_k^- := I_k^- \setminus \{j \in I_k^- : \quad a_j^k > \epsilon\} \ .$$

Calculate

$$\|g^*\| = \min_{g \in \mathrm{co}\{g_j : \ j \in I_k^+ \cup I_k^-\}} \|g\| \ .$$

If $\|g^*\| \leq \eta$ then stop. Else set:

$$\rho_{\min} := \rho_{\min} + r(\rho_{\max} - \rho_{\min}) \tag{1.2.14}$$

and go to step 2.

The main disadvantage of this approach is that two subproblems must be solved at step 3, i.e. (1.2.8) and (1.2.9). We remark that in the objective function of (1.2.8) are present both lower and upper piecewise affine approximations of f. In fact, Δ_p is a convex combination of the local models Δ^+ and Δ^-.

Theorem 1.2.5 [11] *For any $\epsilon > 0$ and $\eta > 0$, DC-NCVX Algorithm stops in a finite number of "main iterations" at a stability center y_k^* satisfying the approximate stationarity condition*

$$\|g^*\| \leq \eta \quad \textit{with } g^* \in \partial_\epsilon^G f(y_k^*) \ . \tag{1.2.15}$$

2 Gradient Sampling Algorithm

We consider the problem (P) in §1. It is well-known that, if the objective function f is smooth, the vector of norm one

$$d^* = \operatorname*{argmin}_{\|d\|\leq 1} f'(x,d)$$

is the so-called normalized "steepest descent" direction of f at x. Coming back to our nonsmooth problem (P), if we assume that f is Lipschitzian near x and regular in the Clarke's sense, from Proposition IV.2.1.1 it follows that d^*, the steepest descent, is the solution of the following minimax problem

$$\min_{\|d\|\leq 1} \max_{g\in\partial f(x)} g^T d \ , \tag{2.0.1}$$

which is practically impossible to solve, since in general it is not provided a description of the whole Clarke gradient $\partial f(x)$.

2.1 The Idea

Before continuing our discussion of the Gradient Sampling algorithm [3], we show the following simple result.

Lemma 2.1.1 [3] *Let C be any compact convex subset of $\mathbb{R}^n$. Then*

(i) $-\operatorname{dist}(0 \mid C) = \min_{\|d\|\leq 1} \max_{g\in C} g^T d$*;*

(ii) *Letting $\hat{g}$ be the least norm element of C, the vector*

$$\hat{d} = -\frac{\hat{g}}{\|\hat{g}\|}$$

solves the problem on the right-hand side of the equality in (i).

Proof. From Sion-Kakutani Theorem (see, e.g., [2]), we have

$$\begin{aligned}-\operatorname{dist}(0 \mid C) &= -\min_{g\in C}\|g\| = -\min_{g\in C}\max_{\|d\|\leq 1} g^T d \\ &= -\max_{\|d\|\leq 1}\min_{g\in C} g^T d = -\max_{\|d\|\leq 1}\min_{g\in C} -g^T d \\ &= \min_{\|d\|\leq 1}\max_{g\in C} g^T d \ .\end{aligned}$$

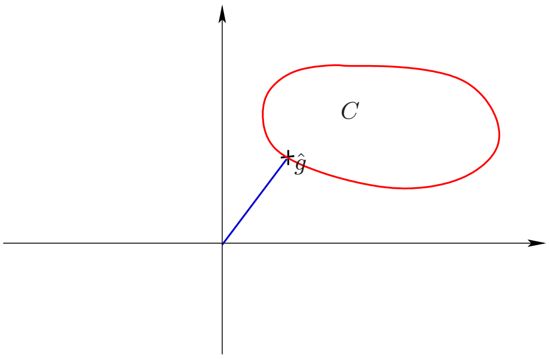


Fig. 2.1.1: $\mathrm{dist}(0 \mid C)$.

Taking into account that $-\|\hat{g}\| = \hat{g}^T\hat{d}$, (ii) follows (see Fig. 2.1.1).

□

Consequently, after finding the minimizer g^* of the problem

$$\min_{g \in \partial f(x)} \|g\| \ , \tag{2.1.1}$$

the normalized steepest descent direction d^* is given by $d^* = -\dfrac{g^*}{\|g^*\|}$.

J. Burke, A. Lewis and L. Overton, in [3], approximate the Clarke gradient of f at x by the “gradient sampling”

$$G_\epsilon(x) \overset{\triangle}{=} \mathrm{co}\{\nabla f(x+d_i) : \quad i = 1, \ldots, m, \ m > n, \quad \text{f is differentiable at } x + d_i\}$$

where, for some small fixed $\epsilon > 0$, $d_1, \ldots, d_m$ is a linearly independent uniformly distributed random collection of vectors of $B_\epsilon^{(n)}(0)$. By solving the following quadratic programming problem

$$\min_{g \in G_\epsilon(x)} \|g\|^2 \ , \tag{2.1.2}$$

and letting $\bar{g}^*$ be the minimizer of above problem, the vector $\bar{d}^* = -\dfrac{\bar{g}^*}{\|\bar{g}^*\|}$ is the approximate normalized steepest descent direction of f at x.

2.2 The Algorithm

At each iteration k, the Gradient Sampling algorithm [3] calculates the approximate steepest descent direction d_k of f at the current iterate x_k and then computes a step t_k along this direction by using a line search procedure. Finally the new iterate is set as

$$x_{k+1} := x_k + t_k d_k \ .$$

Even if the convergence of the method has been proved for locally Lipschitzian functions, the algorithm has been tested on more general functions. In particular, the results of the numerical experimentations appear to be promising for any continuous and almost everywhere differentiable function. A MATLAB implementation of the algorithm is available on the web at the URL http://www.cs.nyu.edu/overton/-papers/gradsamp/alg.

Chapter VI

A New Bundle Method for Nonsmooth Nonconvex Optimization

Introduction. In this chapter we present the new bundle method NonConvexNonSmooth (NCNS) [12], which implements the bundle splitting strategy used for both NCVX and DC-NCVX algorithms and described in §V.1.2. NCNS introduces changes into the local model of the objective function f of the nonsmooth nonconvex problem (P), defined in §V.1, and into the localization of the new "sample point" at k^{th} "main iteration".

1 The Model

We consider the unconstrained minimization problem (V.1.0.1), where the objective function f is assumed to be weakly semismooth as well[1].

We assume that we are able to calculate at each point x both the objective function value $f(x)$ and a subgradient $g \in \partial f(x)$, i.e. an element of the Clarke gradient.

[1] The weak semismoothness assumption implies that f is locally Lipschitzian (see the proof of the Theorem I.1.2.2)

We denote by y_k the current stability center during the execution of the k^{th} "main iteration" and by g_k any subgradient of f at y_k. In the sequel we will assume that a starting point, say x_1, is available and that at least the stability center y_k belongs to the level set defined by x_1, namely

$$y_k \in \mathcal{F}_1 \stackrel{\triangle}{=} \{x \in \mathbb{R}^n : \quad f(x) \leq f(x_1)\}.$$

We assume also that the set $\mathcal{F}_1$ is compact, so we indicate by L_1 a Lipschitz constant of f on $\mathcal{F}_1$. We denote by L_ρ a Lipschitz constant of f on

$$\mathcal{F}_\rho = \left\{ x \in \mathbb{R}^n : \quad \operatorname{dist}(x \mid \mathcal{F}_1) \leq \frac{2L_1}{\rho} \right\}.$$

1.1 The Bundle Splitting Idea

The way to split the bundle is slightly modified with respect to NCVX and DC-NCVX algorithms:

$$I_k^+ \stackrel{\triangle}{=} \{j : \quad \alpha_j^k \geq -\sigma\} \quad \text{and} \quad I_k^- \stackrel{\triangle}{=} \{j : \quad \alpha_j^k < -\sigma\} \ , \tag{1.1.1}$$

for some $\sigma > 0$. We observe that I_k^+ is never empty, as at least the index corresponding to the stability center y_k belongs to the I_k^+.

Let $h(d) \stackrel{\triangle}{=} f(y_k + d) - f(y_k)$ be the difference function. We construct two polyhedral models of h, using separately the two bundles. In particular, setting $\alpha_j^k = \max\{\alpha_j^k, 0\}$, for all $j \in I_k^+$, we define the two piecewise affine functions:

$$\Delta^+(d) \stackrel{\triangle}{=} \max_{j \in I_k^+} \left\{ g_j^T d - \alpha_j^k \right\}$$

and

$$\Delta^-(d) \stackrel{\triangle}{=} \min \left\{ 0, \ \min_{j \in I_k^-} \{ g_j^T d - \alpha_j^k \} \right\},$$

which are convex and concave functions, respectively. We remark that $\Delta^-(d)$ is equal to zero around $d = 0$ and $\Delta^+(0) = \Delta^-(0) = h(0)$.

1.2 The Quadratic Subproblem

Our approach, at current point y_k, consists in finding a tentative stepsize by solving the following problem:

$$d(\rho) = \operatorname*{argmin}_{d \in \mathbb{R}^n} \Delta(d) + \frac{1}{2}\rho\|d\|^2 \tag{1.2.1}$$

where $\Delta \triangleq \Delta^+ - \Delta^-$ and $\rho > 0$ is the proximity parameter introduced for both stabilization and well-posedness purposes (see Fig. 1.2.1). The rationale of the model is in the attempt of locating the new "sample point" so that both the model functions Δ^+ and Δ^- predict reduction, but, at the same time, their predictions are mostly different.

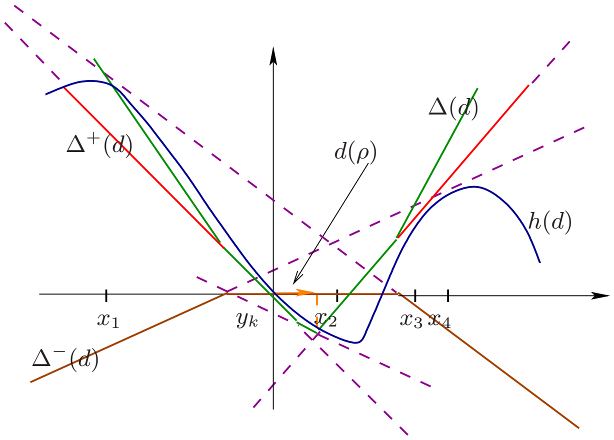


Fig. 1.2.1: NCNS approach.

By introducing the scalar variables v and w, we can rewrite (1.2.1) as the following quadratic programming problem

$$(QP_\rho)\begin{cases} z(\rho) = \min\limits_{v,w,d} v - w + \dfrac{1}{2}\rho\|d\|^2 \\ v \geq g_j^T d - \alpha_j^k \quad j \in I_k^+ \\ w \leq g_j^T d - \alpha_j^k \quad j \in I_k^- \\ w \leq 0 \end{cases}$$

The dual of QP_ρ is

$$(DP_\rho)\begin{cases} \zeta(\rho) = \min\limits_{\lambda\geq 0,\mu\geq 0} \dfrac{1}{2\rho}\|G_+\lambda - G_-\mu\|^2 + \lambda^T\alpha_+^k - \mu^T\alpha_-^k \\ \\ e^T\lambda = 1 \\ e^T\mu \leq 1 \ , \end{cases}$$

where $G_+,\ G_-,\ \alpha_+^k$ and α_-^k are the same quantities we have defined in §V.1.2.

By indicating $(d(\rho), v(\rho), w(\rho))$ and $(\lambda(\rho), \mu(\rho))$ the optimal solutions of (QP_ρ) and (DP_ρ) respectively, the following primal-dual relations hold:

$$d(\rho) = \frac{1}{\rho}\left(-G_+\lambda(\rho) + G_-\mu(\rho)\right) \tag{1.2.2a}$$

$$v(\rho) = d(\rho)^T G_+\lambda(\rho) - \lambda(\rho)^T\alpha_+^k \tag{1.2.2b}$$

$$w(\rho) = d(\rho)^T G_-\mu(\rho) - \mu(\rho)^T\alpha_-^k \tag{1.2.2c}$$

We observe that, since the triple $(d, v, w) = (0, 0, 0)$ is feasible for (QP_ρ), $z(\rho) \leq 0$. Consequently, $v(\rho) \leq -\frac{1}{2}\rho\|d(\rho)\|^2 + w(\rho) \leq 0$ and $v(\rho) - w(\rho) \leq 0$.

Our approach solves only one quadratic subproblem, that is (QP_ρ), and a kind of trust region is implicity defined, i.e.

$$\mathcal{C} \triangleq \{d:\ \Delta^+(d) \leq \Delta^-(d)\ \},$$

as it happens in NCVX.

The function Δ participating in the objective function of subproblem (1.2.1) contains both lower and upper approximations Δ^+ and Δ^-, such as Δ_p participating in the subproblem (V.1.2.8) for DC-NCVX.

1.3 Some Results

Before giving a formal description of the algorithm, we state some simple properties of problem (QP_ρ), assuming that ρ is not smaller than a fixed positive threshold $\rho_{\min}$.

Lemma 1.3.1 *Let $\delta > 0$ and $v(\rho) - w(\rho) \geq -\delta$, then*

$$\frac{1}{\rho}\|G_+\lambda(\rho) - G_-\mu(\rho)\|^2 \leq \delta.$$

Proof. The property follows by (1.2.2) and by, respectively, nonnegativity of α_+^k and negativity of α_-^k. In fact we have

$$-\delta \le v(\rho) - w(\rho) = -\frac{1}{\rho}\|G_+\lambda(\rho) - G_-\mu(\rho)\|^2 - \lambda(\rho)^T\alpha_+^k + \mu(\rho)^T\alpha_-^k$$

$$\le -\frac{1}{\rho}\|G_+\lambda(\rho) - G_-\mu(\rho)\|^2.$$

□

Lemma 1.3.2 *For any $\rho > 0$ the following inequality holds:*

$$\|d(\rho)\| \le \frac{2L_1}{\rho}.$$

Proof. Since $\frac{1}{2}\rho\|d(\rho)\|^2 + v(\rho) - w(\rho) \le 0$, we have

$$v(\rho) + \frac{1}{2}\rho\|d(\rho)\|^2 \le w(\rho) \le 0.$$

Hence, taking into account that

$$0 \ge v(\rho) + \frac{1}{2}\rho\|d(\rho)\|^2 \ge g_k^T d(\rho) + \frac{1}{2}\rho\|d(\rho)\|^2 \ge -\|g_k\|\|d(\rho)\| + \frac{1}{2}\rho\|d(\rho)\|^2,$$

the thesis follows by considering that $\|g_k\| \le L_1$ as by hypothesis y_k belongs to $\mathcal{F}_1$.

□

Lemma 1.3.3 *Let $\hat{\alpha_\rho} \overset{\triangle}{=} f(y_k) - f(y_k + d(\rho)) + \hat{g}^T d(\rho)$, with $\hat{g} \in \partial f(y_k + d(\rho))$. For $\sigma > 0$ there exists a threshold value $\bar{\rho}(\sigma)$ such that $|\hat{\alpha_\rho}| \le \sigma$ for all $\rho \ge \bar{\rho}(\sigma)$.*

Proof. By Lemma 1.3.2, we have, for all $\rho > 0$, $\|d(\rho)\| \le \dfrac{2L_1}{\rho}$. Since $y_k + d(\rho) \in \mathcal{F}_\rho$, taking into account $\rho \ge \rho_{\min}$, we have

$$|\hat{\alpha_\rho}| \le |f(y_k) - f(y_k + d(\rho))| + |\hat{g}^T d(\rho)| \le 2L_{\rho_{\min}}\|d(\rho)\| \le \frac{4L_{\rho_{\min}}L_1}{\rho}.$$

Consequently for the threshold value we have $\bar{\rho}(\sigma) = \dfrac{4L_{\rho_{\min}}L_1}{\sigma}$.

□

2 The Algorithm

The scheme of the entire algorithm coincides with Algorithm V.1.2.1 and the following global parameters are to be set:

- the stationarity tolerance $\eta > 0$ and the distance parameter $\epsilon > 0$;
- the descent parameter $m_1 \in (0,1)$ and the concave cut parameter $m_2 > 1$;
- the lower threshold on the proximity parameter $\rho_{\min} > 0$;
- the increase parameter $R > 1$;
- the linearization error threshold parameter $\sigma > 0$;

The initialization of the algorithm requires a starting point $x_1 \in \mathbb{R}^n$. The initial stability center is set equal to x_1. The initial bundle is made up by just one element $(x_1, f(x_1), g_1, 0, 0)$, where $g_1 \in \partial f(x_1)$. Consequently I_1^- is the empty set, while I_1^+ is a singleton.

Now we can describe our "main iteration".

Algorithm 2.0.1 (k^{th} NCNS *main iteration*)

Step 0. *Select* $\rho \geq \rho_{\min}$.

Step 1. *Solve either* (QP_ρ) *or* (DP_ρ) *and determine* $(d(\rho), v(\rho), w(\rho))$ *and* $(\lambda(\rho), \mu(\rho))$.
Set $\delta = \dfrac{\eta^2}{\rho}$. *If* $v(\rho) - w(\rho) < -\delta$ *go to step 3.*

Step 2. *Two cases can occur:*

a) $\mu(\rho) = 0$. *If* $a_j^k \leq \epsilon$ *for all* $j \in I_k^+$ *such that* $\lambda_j(\rho) > 0$ *then stop, else delete from* I_k^+ *all the j's such that* $a_j^k > \epsilon$, *set* $\rho = R\rho$ *and return to step* 1*;*

b) $\mu(\rho) \neq 0$. *Reset* I_k^-. *Set* $\rho = R\rho$ *and return to step* 1.

Step 3. *Evaluate* $f(y_k + d(\rho))$. *If* $f(y_k + d(\rho)) - f(y_k) \leq m_1 v(\rho)$, *then exit from main iteration (stability center update).*

Step 4. *Calculate* $\hat{g} \in \partial f(y_k + d(\rho))$, *and*[2] $\hat{\alpha} \overset{\triangle}{=} f(y_k) - f(y_k + d(\rho)) + \hat{g}^T d(\rho)$.

Step 5. *Three cases can occur:*

a) $\hat{\alpha} \geq 0$. *Update the bundle, by introducing a new element in* I_k^+, *and return to step* 1*;*

[2] To simply the notation, we drop the subscript ρ for $\hat{\alpha}_\rho$.

b) $-\sigma \leq \hat{\alpha} < 0$. *Set* $\hat{\alpha} = 0$ *and find a scalar* $t \in (0, 1]$ *such that* $g(t) \in \partial f(y_k + td(\rho))$ *satisfies the condition* $g(t)^T d(\rho) \geq m_1 v(\rho)$. *Update the bundle, by introducing a new element in* I_k^+*, and return to step* 1*;*

c) $\hat{\alpha} < -\sigma$. *If* $\hat{g}^T d(\rho) - \hat{\alpha} \leq m_2 w(\rho)$*, update the bundle by introducing a new element in* I_k^-. *Set* $\rho = R\rho$ *and return to step* 1.

2.1 Some Remarks

Before discussing the convergence, the following remarks are in order.

Remark 2.1.1 *We observe that from the definition of* $\hat{\alpha}$*, whenever cases a) and b) at step* 5 *of the main iteration occur, the condition* $\hat{g}^T d(\rho) - \hat{\alpha} > m_1 v(\rho)$ *holds. On the other hand whenever case b) occurs, since* $\hat{\alpha}$ *is set equal to zero, the condition* $\hat{g}^T d(\rho) - \hat{\alpha} > m_1 v(\rho)$ *holds as well. We observe that the problem of finding the scalar t, at case b), is well-posed. In fact since the directional derivative* $f'(y_k + td(\rho), d(\rho))$ *exists for any* $t \geq 0$*, from the mean value theorem (Theorem IV.3.1.3) it follows that*

$$f(y_k + d(\rho)) - f(y_k) = c \tag{2.1.1}$$

for some $c \in [f'_{\inf}, f'_{\sup}]$*, where*

$$f'_{\inf} \stackrel{\triangle}{=} \inf_{0 \leq t \leq 1} f'(y_k + td(\rho), d(\rho)) \quad \textit{and} \quad f'_{\sup} \stackrel{\triangle}{=} \sup_{0 \leq t \leq 1} f'(y_k + td(\rho), d(\rho))$$

Moreover, taking into account that the sufficient decrease condition is not satisfied, i.e.

$$f(y_k + d(\rho)) - f(y_k) > m_1 v(\rho),$$

by (2.1.1) and the definition of $f'_{\sup}$ *there exists a scalar* $\bar{t} \in (0, 1]$ *such that*

$$f'(y_k + \bar{t}d(\rho), d(\rho)) > m_1 v(\rho)$$

and, by weak semismoothness assumption, we have

$$\lim_{t \downarrow \bar{t}} g(t)^T d(\rho) > m_1 v(\rho),$$

where $g(t) \in \partial f(y_k + td(\rho))$. *Consequently the inequality* $g(t)^T d(\rho) > m_1 v(\rho)$ *holds in some interval* $(\bar{t}, \hat{t})$.

Remark 2.1.2 *When the stopping criterion at step 2 is met, the following conditions hold:*

$$v(\rho) - w(\rho) \geq -\delta; \tag{2.1.2}$$

$$\mu(\rho) = 0; \tag{2.1.3}$$

$$a_j^k \leq \epsilon \quad \forall\, j:\ \lambda_j(\rho) > 0. \tag{2.1.4}$$

Hence, by equations (1.2.2), we have

$$-\delta \leq v(\rho) - w(\rho) = -\frac{1}{\rho}\|G_+\lambda(\rho) - G_-\mu(\rho)\|^2 - \lambda(\rho)^T\alpha_+^k + \mu(\rho)^T\alpha_-^k$$

From (2.1.3) and nonnegativity of α_+^k, we have

$$\frac{1}{\rho}\|G_+\lambda(\rho)\|^2 \leq \delta \tag{2.1.5}$$

which in turn, taking into account (2.1.4) and $\delta \leq \dfrac{\eta^2}{\rho}$, implies both $g^* \triangleq G_+\lambda(\rho) \in \partial_\epsilon^G f(y_k^*)$ and $\|g^*\| \leq \eta$.

3 Convergence

In this section we prove the termination of the algorithm at a point satisfying an approximate stationarity condition. In particular we prove that, for any given $\epsilon > 0$ and $\eta > 0$, it is possible to set the input parameters so that, after a finite number of "main iteration" executions, the algorithm stops at a point y_k^* satisfying the condition

$$\|g^*\| \leq \eta \text{ ,with } g^* \in \partial_\epsilon^G f(y_k^*)\ .$$

3.1 Termination of the "Main Iteration"

We start proving the following:

Lemma 3.1.1 *The k^{th} "main iteration" algorithm cannot loop infinitely many times without entering step 2.*

Proof. Suppose that the algorithm loops infinitely many times (i.e. the descent test at step 3 is never satisfied) without entering step 2. We index by i all the quantities referred to the i^{th} passage through steps 1-5.

We observe that the case c) at step 5 cannot occur infinitely many times. In fact, whenever $\hat{\alpha}_i < -\sigma$, the parameter ρ_i is increased and consequently there exists an index $\hat{\imath}$ such that $\hat{\alpha}_i \geq -\sigma$ for all $i > \hat{\imath}$ (see Lemma 1.3.3). Thus only cases a) or b) can occur infinitely many times and, taking into account Remark 2.1.1, the condition $d_i(\rho_i)^T \hat{g}_i - \hat{\alpha}_i > m_1 v_i(\rho_i)$ is met infinitely many times too. Letting $\bar{\imath}$ index the last passage through case c) at step 5, we note that ρ_i remains constant for all $i \geq \bar{\imath}$. Consequently the sequence $\{z_i(\rho_i)\}$, for $i \geq \bar{\imath}$, is monotonically nondecreasing, bounded from above, and hence convergent. Moreover, the condition $\rho_i \geq \rho_{\min}$ implies, by Lemma 1.3.2, $\|d_i(\rho_i)\| \leq \dfrac{2L_1}{\rho_{\min}}$, hence $\{d_i(\rho_i)\}$ belongs to a compact set and there exists a convergent subsequence, say $\{d_i(\rho_i)\}_{i\in I'}$. Thus the subsequence $\{v_i(\rho_i) - w_i(\rho_i)\}_{i\in I'}$ is convergent. Furthermore, we have

$$-\delta > v_i(\rho_i) \geq g_k^T d_i(\rho_i) \geq -\frac{2L_1^2}{\rho_{\min}},$$

hence $\{v_i(\rho_i)\}_{i\in I'}$ is bounded and consequently there exist two convergent subsequences $\{v_i(\rho_i)\}_{i\in I''\subset I'}$ and $\{w_i(\rho_i)\}_{i\in I''\subset I'}$. Now let t and s be two successive indices in I'' and let $\bar{v} = \lim_{i\in I''} v_i(\rho_i)$, then we have:

$$\begin{aligned} d_t(\rho)^T \hat{g}_t - \hat{\alpha}_t &\geq m_1 v_t(\rho_t) \\ d_s(\rho)^T \hat{g}_t - \hat{\alpha}_t &\leq v_s(\rho_s), \end{aligned}$$

that is $v_s(\rho_s) - m_1 v_t(\rho_t) \geq (d_s(\rho_s) - d_t(\rho_t))^T \hat{g}_t$, which implies $\bar{v} \geq 0$. Observe that $\bar{v} \geq 0$ contradicts the hypothesis that the algorithm never enters step 2. In fact, in this case, we would have:

$$v_i(\rho_i) < w_i(\rho_i) - \delta < -\delta \quad \forall\, i,$$

which, taking into account $v_i(\rho_i) \to \bar{v}$, would imply $\bar{v} \leq -\delta$.

□

Now we can prove finite termination of the k^{th} “main iteration”.

Lemma 3.1.2 *If the k^{th} "main iteration" does not terminate with satisfaction of the sufficient decrease condition at step 3, then the stopping condition is met after a finite number of passages through step 2.*

Proof. Assume that the sufficient decrease condition at step 3 of the "main iteration" is never satisfied. If we assume also that the stopping condition is never satisfied, by Lemma 3.1.1 we have that step 2 is entered infinitely many times and the parameter ρ grows indefinitely.

Let $\tilde{\imath}$ be an index such that $\rho_{\tilde{\imath}} \geq \dfrac{2L_1}{\epsilon}$ and, consequently, all points newly generated by the algorithm for $i \geq \tilde{\imath}$ are characterized by a distance from the stability center $a_j^k \leq \epsilon$ (Lemma 1.3.2). We observe also that (see the proof of Lemma 3.1.1), for sufficiently large values of ρ, only modifications of I_k^+ can occur. Thus, taking into account the reset of I_k^- and the deletion of elements of I_k^+, for sufficiently large values of ρ, we have $I_k^- = \emptyset$ and $a_j^k \leq \epsilon$ for all $j \in I_k^+$. Thus, the stopping condition is met after a finite number of passages through step 2.

□

Remark 3.1.1 *The proofs of the previous lemmas ensure also that the value of the proximity parameter ρ cannot become arbitrarily large.*

3.2 Convergence of the Algorithm

Now we are ready to prove the overall finiteness of the algorithm.

Theorem 3.2.1 *For any $\epsilon > 0$ and $\eta > 0$, the algorithm stops in a finite number of "main iterations" at a stability center y_k^* satisfying the approximate stationarity condition*

$$\|g^*\| \leq \eta, \qquad \text{with } g^* \in \partial_\epsilon^G f(y_k^*). \tag{3.2.1}$$

Proof. We prove, by contradiction, that the stopping criterion is satisfied in a finite number of "main iterations". Suppose in fact that an infinite number of main iterations occurs. Then the descent condition at step 3 is verified infinitely many times. Let y_k be the stability center at the k^{th} main iteration and $v^{(k)}$ and $\delta^{(k)}$ be, respectively, the values of $v(\rho)$ and δ for which the descent condition at step 3 has been fulfilled. Then

$$f(y_{k+1}) \leq f(y_k) + m_1 v^{(k)}$$

and, after k main iterations,

$$f(y_k) \leq f(y_1) - km_1\delta^{(k)} \leq f(y_1) - km_1\frac{\eta^2}{\rho_{\max}},$$

where ρ_{max} is any upper bound on the proximity parameter ρ (see Remark 3.1.1) Passing to the limit we have

$$\lim_{k\to\infty} f(y_k) - f(y_1) \leq -\infty,$$

which is a contradiction, since f is bounded from below by hypothesis.

□

4 Implementation

It is worth noting that the algorithm described in §2 can produce a bundle whose size can grow indefinitely. Thus, to make the method implementable, it is important to introduce bounded storage for the bundle. Of course it is necessary as well to show that convergence properties proved in §3 are retained under such hypothesis.

To tackle the problem we introduce an aggregation technique scheme (see Chapter II). In particular let $(d(\rho), v(\rho), w(\rho))$ and $(\lambda(\rho), \mu(\rho))$ be, respectively, the solution of (QP_ρ) and (DP_ρ) at step 1 of the "main iteration". If we define the aggregate quantities

$$g_p \triangleq G_+\lambda(\rho)\ ,\ \alpha_p \triangleq \lambda(\rho)^T\alpha_+^k$$

and, in case $\mu(\rho) \neq 0$,

$$g_m \triangleq G_-\mu(\rho)\ ,\ \alpha_m \triangleq \mu(\rho)^T\alpha_-^k,$$

it is easy to verify that the aggregate problem QP_ρ^a

$$QP_\rho^a \begin{cases} \min\limits_{v,w,d} & \dfrac{1}{2}\rho\|d\|^2 + v - w \\ & v \geq g_p^T d - \alpha_p \\ & v \geq g_j^T d - \alpha_j^k \qquad j \in \bar{I}_k^+ \\ & w \leq g_m^T d - \alpha_m \\ & w \leq g_j^T d - \alpha_j^k \qquad j \in \bar{I}_k^- \end{cases}$$

has the same optimal solution $(d(\rho), v(\rho), w(\rho))$ as (QP_ρ), where $\bar{I}_k^+$ and $\bar{I}_k^-$ are arbitrary subsets of I_k^+ and I_k^- respectively. Of course, in case $I_k^- = \emptyset$ or $\mu(\rho) = 0$, the formulation of the aggregate problem does not contain the constraint $w \leq g_m^T d - \alpha_m$ and $(d(\rho), v(\rho), 0)$ is still optimal.

Now suppose that at a certain execution of the "main iteration", the quadratic program (QP_ρ) (or (DP_ρ)) is solved, and the corresponding optimal dual vector $(\lambda(\rho), \mu(\rho))$ is calculated. If we calculate also the quantities g_p, α_p, g_m, α_m, it is possible to construct the aggregate problem (QP_ρ^a) by inserting the aggregated constraints into (QP_ρ) and deleting part of its bundle elements. Thus, the new quadratic program can be obtained by inserting the new constraint, corresponding to the new bundle element calculated at step 5 of the "main iteration", into the aggregated problem (QP_ρ^a). Of course, such an aggregation task will only be carried out each time a given maximal bundle dimension is reached.

The aggregation mechanism does not impair convergence. Indeed the key argument is that the monotonicity of the sequence $\{z_i(\rho_i)\}$, necessary in the proof of Lemma 3.1.1, is still guaranteed.

The algorithm, equipped with the aggregation scheme, has been implemented in double precision Fortran-77 under the *Windows XP* system.

5 Numerical Examples

Our code, called NCNS, has been tested on a set of 25 problems [20] available on the web at the URL http://www.cs.cas.cz/~luksan/test.html. All test problems, except the Rosenbrock problem, are nonsmooth.

The input parameters have been set as follows: $\epsilon = 10^{-2}$, $\eta = 10^{-4}$, $\sigma = 10^{-2}$, $m_1 = 0.2$, $m_2 = 1.2$, $\rho_{min} = 10^{-2}\|g(x_1)\|$, $R = 6$. In table 5.0.1 we report the computational results in terms of the number N_f of function evaluations. By f^* and f we indicate, respectively, the minimum value of the objective function and the function value reached by the algorithm when the stopping criterion is met.

At each iteration we solve the dual program (DP_ρ), by using the subroutine DQPROG provided by the IMSL library and based on M.J.D. Powell's implementation of the Goldfarb and Idnani [13] dual quadratic programming algorithm.

We compare the results provided by our code NCNS with NCVX [10] and the variable metric algorithms VN [21] and VMNC [41]. The performance of our algorithm seems comparable with those of the considered methods.

Problem				NCNS		NCVX		VN		VMNC	
#	Problem	n	f^*	N_f	f	N_f	f	N_f	f	N_f	f
1	Rosenbrock	2	0	54	5.137e-06	70	5.009e-07	34	2.759e-11	33	0.320e-07
2	Crescent	2	0	53	5.112e-06	22	8.022e-06	16	9.489e-11	15	0.949e-10
3	CB2	2	1.9522245	16	1.9522255	18	1.9522245	17	1.9522247	16	1.9522250
4	CB3	2	2	14	2.0000000	15	2.0000000	17	2.0000000	17	2.0000000
5	DEM	2	-3	13	-3.0000000	21	-2.9999999	20	-2.9999996	20	-2.9999997
6	QL	2	7.2	15	7.2000001	28	7.2000005	19	7.2000000	18	7.2000023
7	LQ	2	-1.4142136	15	-1.4142136	9	-1.4142135	10	-1.4142133	10	-1.4142133
8	Mifflin1	2	-1	165	-9.9993895	127	-0.9999977	127	-0.9999924	59	-0.9999925
9	Mifflin2	2	-1	14	-9.9999915	13	-1.0000000	13	-0.9999997	35	-0.9999998
10	Rosen-Suzuki	4	-44	33	-43.999997	29	-44.000000	38	-43.999999	32	-43.999975
11	Shor	5	22.600162	27	22.600163	44	22.600162	40	22.600162	30	22.600186
12	Maxquad	10	-0.8414083	90	-0.8413860	56	-0.8414078	89	-0.8414057	89	-0.8414057
13	Maxq	20	0	187	1.561e-06	293	1.660e-07	123	1.468e-06	111	0.898e-05
14	Maxl	20	0	23	4.493e-15	44	1.110e-15	23	0.0000000	23	0.0000000
15	Goffin	50	0	56	1.984e-13	148	1.142e-13	360	4.153e-06	368	0.332e-05
16	El-Attar	6	0.5598131	172	0.5598143	152	0.5598163	83	0.5598155	76	0.5598184
17	Wolfe	2	-8	43	-7.9999998	21	-7.9999998	14	-7.9999998	14	-7.9999998
18	MXHILB	50	0	24	1.764e-05	33	1.768e-05	66	3.272e-06	67	0.201e-05
19	L1HILB	50	0	30	1.709e-05	104	6.978e-07	67	9.457e-07	64	0.153e-05
20	Colville1	5	-32.348679	36	-32.348677	47	-32.348679	53	-32.348678	47	-32.348675
21	Gill	10	9.7857721	308	9.7858381	164	9.7857746	241	9.7858732	108	9.7862324
22	HS78	5	-2.9197004	237	-2.9191783	159	-2.9196589	32	-2.9197003	—	———
23	TR48	48	-638565	1662	-638514.80	353	-638565.00	359	-638564.91	295	-638562.27
24	Shell Dual	15	32.348679	642	32.348687	1497	32.349404	315	32.349159	289	32.349018
25	Steiner2	12	16.703838	96	16.703844	196	16.703838	89	16.703838	62	16.703937

Table 5.0.1: NCNS: computational results.

Appendix: Background Material

A - Linear Algebra and Set Theory

Definition A.1 (Linear subspace)[2] *A linear subspace in $\mathbb{R}^n$ is a nonempty subset of $\mathbb{R}^n$ which is closed with respect to addition of vectors and multiplication by reals.*

A linear combination of elements $x_1, \dots, x_k$ in $\mathbb{R}^n$ is a element of the form

$$\sum_{i=1}^{k} \lambda_i x_i \ , \quad \lambda_i \in \mathbb{R} \ .$$

A collection $x_1, \dots, x_k$ of n-dimensional vectors is called linearly independent, if no nontrivial linear combination of the vectors is zero. A collection $x_1, \dots, x_k$ linearly independent is called a basis of a linear subspace C and k is said to be the dimension of C, if every vector from C is a linear combination of the vectors $x_1, \dots, x_k$.

Definition A.2 (Affine subspace)[2] *An affine subspace in $\mathbb{R}^n$ is a set of the form*

$$C = a + B = \{a + x : \quad x \in B\} \ ,$$

where B is a linear subspace in $\mathbb{R}^n$ and a is a vector from $\mathbb{R}^n$.

An affine combination of elements $x_1, \dots, x_k$ in $\mathbb{R}^n$ is a element of the form

$$\sum_{i=1}^{k} \lambda_i x_i \ ,$$

where the coefficients λ_i satisfy $\sum_{i=1}^{k} \lambda_i = 1$.

A collection $x_1, \dots, x_k$ of n-dimensional vectors is called affinely independent, if no nontrivial linear combination of the vectors with zero sum of coefficients is zero, i.e.

$$\sum_{i=1}^{k} \lambda_i x_i = 0, \quad \sum_{i=1}^{k} \lambda_i = 0$$
$$\Rightarrow \lambda_i = 0, \ \forall \ i = 1, \dots, k$$

The affine dimension of an affine subspace $C = a + B$ is the dimension of the linear subspace B.

Definition A.3 (Affine hull) [2] *Let C a subset of $\mathbb{R}^n$. The affine hull of a set C, denoted as* $\operatorname{aff} C$*, is the smallest affine subspace containing C.*

A convex combination of elements $x_1, \dots, x_k$ in $\mathbb{R}^n$ is a element of the form

$$\sum_{i=1}^{k} \lambda_i x_i \ ,$$

where the coefficients λ_i satisfy $\sum_{i=1}^{k} \lambda_i = 1$ and $\lambda_i \geq 0$, $i = 1, \dots, k$.

Definition A.4 (Convex set) [35] *A subset C of $\mathbb{R}^n$ is said to be convex if*

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

whenever $x, y \in C$ and $\lambda \in (0, 1)$.

Definition A.5 (Convex hull) [2] *Let C a subset of $\mathbb{R}^n$. The convex hull of the set C, denoted as* $\operatorname{co} C$*, is the smallest convex set containing C.*

A set $C \subset \mathbb{R}^n$ is called closed, it it contains limits of all converging sequences of elements of C, namely

$$\left\{ x_i \in C, \quad x = \lim_{i \to \infty} x_i \right\}$$
$$\Rightarrow x \in C \ .$$

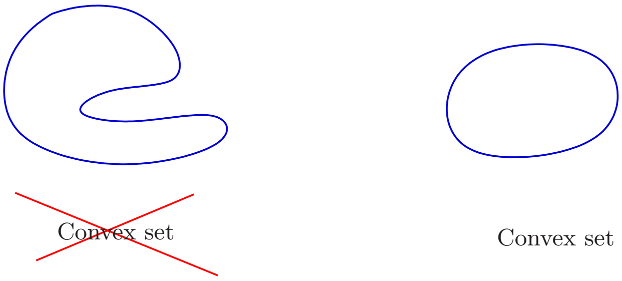


Fig. A.1: Convex sets.

Definition A.6 (Closure) [2] *Let C a nonempty subset of $\mathbb{R}^n$. The closure of set C, denoted as $\operatorname{cl} C$, is the smallest closed set containing C.*

The closed convex hull of a nonempty set $C \subset \mathbb{R}^n$, denoted as $\overline{\text{co}}\ C$, is the intersection of all closed convex set containing C.

A set is called countable if it is the range of some sequence and finite if it is the range of some finite sequences (see, e.g., [14]).

Definition A.7 (Dense sets) [36] *C is a dense subset of $\mathbb{R}^n$, if for every $\epsilon > 0$, for every $x \in \mathbb{R}^n$ there exists $y \in C$ such that $\|x - y\| < \epsilon$. If $\operatorname{cl} C = \mathbb{R}^n$, C is a dense subset of $\mathbb{R}^n$.*

Definition A.8 (Interior) [2] *Let C a nonempty subset of $\mathbb{R}^n$. A point $x \in C$ is an interior point for C, if some neighborhood of the point is contained in C, i.e. there exists $r > 0$ such that*

$$B_r^{(n)}(x) \subset C \ ,$$

where $B_r^{(n)}(x)$ is the ball of radius r centered at x.

The interior of C, indicated as $\operatorname{int} C$, is the set of all interior points of C.

Definition A.9 (Relative interior) [2] *Let C a nonempty subset of $\mathbb{R}^n$. A point $x \in C$ is a relative interior for C, if C contains the intersection of a small enough ball centered at x with $\operatorname{aff} C$, i.e. there exists $r > 0$ such that*

$$U_r(\bar{x}) \triangleq B_r^{(n)}(x) \cap \operatorname{aff} C \subset C \ .$$

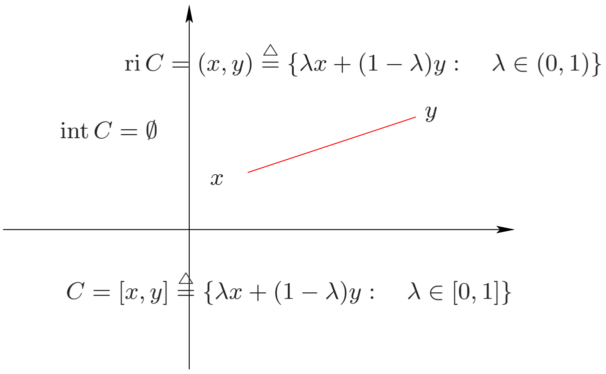

$\operatorname{ri} C = (x, y) \triangleq \{\lambda x + (1-\lambda)y : \quad \lambda \in (0,1)\}$
$\operatorname{int} C = \emptyset$
y
x
$C = [x, y] \triangleq \{\lambda x + (1-\lambda)y : \quad \lambda \in [0,1]\}$

Fig. A.2: Relative interior.

The relative interior of C, indicated as $\operatorname{ri} C$, is the set of all relative interior points of C.

If affine hull of C coincides with $\mathbb{R}^n$, then the relative interior of C is the interior of C. Else $\operatorname{int} C = \emptyset$.

Finally we define the boundary $\partial\, C$ (or relative boundary $\partial_{\rm ri} C$) of C the set

$$\operatorname{cl} C \setminus \operatorname{int} C \quad (\text{or } \operatorname{cl} \mathrm{C} \setminus \operatorname{ri} \mathrm{C}) \ .$$

B - Measure Theory

Definition B.1 (σ-Algebra) [14] *Let S be the whole space. An σ-algebra $\mathcal{C}$ is a class of subsets of S such that*

1. $S, \emptyset \in \mathcal{C}$,
2. $\{C_n\}_{n\in\mathbb{N}}$, $C_n \in \mathcal{C} \quad \Rightarrow \bigcup_{n=1}^{\infty} C_n \in \mathcal{C}$,
3. $C \in \mathcal{C} \Rightarrow S \setminus C \in \mathcal{C}$.

A rectangle in $\mathbb{R}^n$ is a set of the form

$$R \triangleq \{x \in \mathbb{R}^n : \quad a_i < x_i < b_i,\ i = 1, \ldots, n\}$$

and its volume $v(R)$ is given by

$$(b_1 - a_1) \cdots (b_n - a_n)$$

Definition B.2 (Lebesgue measure) [40] *The Lebesgue measure of a set $C \subset \mathbb{R}^n$ is the number*

$$\lambda_n(C) \triangleq \inf \sum_k v(R_k) \ ,$$

where $\{R_k\}$ is a sequence of rectangles covering C and the inf *is taken over all.*

Let λ_n be the Lebesgue measure. Then a set $C \subset \mathbb{R}^n$ is called Lebesgue measurable, if for every set $B \subset \mathbb{R}^n$ we have

$$\lambda_n(B) = \lambda_n(B \cap C) + \lambda_n(B \setminus C)$$

and the class of Lebesgue measurable sets is a σ-Algebra.

There exist nonmeasurable sets (see, e.g., [36]) in the sense of Lebesgue measure.

Definition B.3 (Measure) [40] *Let $\mathcal{C}$ be a σ-algebra and let $\mu : \mathcal{C} \to \mathbb{R}_+ \cup \{+\infty\}$ be a set function. Then μ is a measure, if it is countably additive and such that $\mu(\emptyset) = 0$, i.e.*

$$\mu(\emptyset) = 0 \tag{B.1a}$$

$$\{C_n\}_{n \in \mathbb{N}} \in \mathcal{C}, \quad C_i \cap C_j = \emptyset \ i \neq j \quad \Rightarrow \quad \mu(\bigcup_{n=1}^{\infty} C_n) = \sum_{n=1}^{\infty} \mu(C_n) \tag{B.1b}$$

Let $\mathcal{L}_n$ be the class of Lebesgue measurable sets. Then Lebesgue measure λ_n is a measure on the measurable space $(\mathbb{R}^n$,$\mathcal{L}_n)$.

Definition B.4 (Measure space) [14] *A triple $(S, \mathcal{C}, \mu)$, where S is the space, $\mathcal{C}$ is a σ-algebra over S, and μ is a measure on the measurable space (S,$\mathcal{C}$), is called a measure space.*

Definition B.5 (Lebesgue measurable functions) [40] *Consider the measure space $(\mathbb{R}^n, \mathcal{L}_n, \lambda_n)$. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be Lebesgue measurable, if for every open interval I the set*

$$\{x \in \mathbb{R}^n : \quad y = f(x), \ y \in I\}$$

is Lebesgue measurable.

C - Lipschitzian Functions

The Lipschitz property is largely used in this thesis.

Definition C.1 (Lipschitzian functions)[40] *Let f be a real-valued function defined on a set $C \subset \mathbb{R}^n$. The function f is said to be Lipschitzian on C, if there exists a positive number L such that*

$$|f(x) - f(y)| \leq L\|x - y\| \quad \forall\ x, y \in C$$

The following class of functions is very important in nonsmooth optimization.

Definition C.2 [5] *Let f be a real-valued function defined on an open set $C \subset \mathbb{R}^n$, and let x be a point of C. The function f is said to be Lipschitzian near x, if there exist a positive number L and a small number $\epsilon > 0$ such that*

$$|f(y) - f(z)| \leq L\|y - z\| \quad \forall\ y, z \in x + \epsilon B^{(n)},$$

where $B^{(n)}$ is the unit ball.

Now we define the notion of locally Lipschitzian functions.

Definition C.3 (Locally Lipschitzian functions)[23] *Let f be a real-valued function defined on $\mathbb{R}^n$ and Lipschitzian on each bounded subset of $\mathbb{R}^n$. Then f is called locally Lipschitzian.*

Bibliography

[1] G.T. Bagni. Differenziale e infinitesimo alle origini del Calcolo infinitesimale: note storiche ed esperienze didattiche. In *Atti del convegno per i sessantacinque anni di Francesco Speranza*, pages 12–17, Bologna, 1997.

[2] A. Ben-Tal and A. Nemirovsvki. *Optimization I-II.* 2004. Lectures notes.

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

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

[5] F.H. Clarke. *Optimization and nonsmooth analysis.* John Wiley and Sons, 1983.

[6] V. F. Demyanov and A. Rubinov. *Quasidifferential calculus.* Optimization Software Inc., New York, 1986.

[7] V. F. Demyanov and A. Rubinov. *Constructive nonsmooth analysis.* Verlag Peter Lang, 1995.

[8] A. Fuduli. *Metodi numerici per la minimizzazione di funzioni convesse nondifferenziabili.* PhD thesis, Università della Calabria, Italy, 1997.

[9] A. Fuduli and M. Gaudioso. Tuning strategy for the proximity parameter in convex minimization. *Journal of Optimization Theory and Applications*, 130(1):95–112, 2006.

[10] A. Fuduli, M. Gaudioso, and G. Giallombardo. Minimizing nonconvex nonsmooth functions via cutting planes and proximity control. *SIAM Journal on Optimization*, 14:743–756, 2004.

[11] A. Fuduli, M. Gaudioso, and G. Giallombardo. A DC piecewise affine model and a bundling technique in nonconvex nonsmooth minimization. *Optimization Methods and Software*, 19:89–102, 2004.

[12] M. Gaudioso, E. Gorgone, and M.F. Monaco. Piecewise linear approximations in nonconvex nonsmooth optimization. Technical Report 6/06, Laboratorio di Logistica, Dipartimento di Elettronica Informatica e Sistemistica, Università della Calabria, Italy, 2006.

[13] D. Goldfarb and A. Idnani. A numerically stable dual method for solving strictly convex quadratic program. *Mathematical Programming*, 27:1–33, 1983.

[14] P. R. Halmos. *Measure Theory.* Spinger-Verlag, New York, 1970.

[15] J.B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms Vol. I.* Springer-Verlag, 1993.

[16] J.B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms Vol. II.* Springer-Verlag, 1993.

[17] J. E. Kelley. The cutting-plane method for solving convex programs. *Journal of SIAM*, 8(4):703–712, 1960.

[18] C Lemaréchal and M. Bancora Imbert. Le module M1FC1. Technical Report B.P.,105, INRIA, 1985.

[19] C. Lemaréchal, A. Nemirovskii, and Y. Nesterov. New variants of bundle methods. *Mathematical Programming*, 69(1):111–147, 1995.

[20] L. Lukšan and J. Vlček. Test problems for nonsmooth unconstrained and linearly constrained optimization. Technical Report 798, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, 2000.

[21] L. Lukšan and J. Vlček. Variable Metric Methods for Nonsmooth Optimization. Technical Report 837, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, 2001.

[22] M. Mäkelä and P. Neittaanmäki. *Nonsmooth optimization - Analyisis and algorithms with applications to optimal control.* World Scientific, 1992.

[23] R. Mifflin. An algorithm for constrained optimization with semismooth functions. *Mathematics of Operations Research*, 2(2):191–207, 1977.

[24] R. Mifflin. Semismooth and semiconvex functions in constrained optimization. *SIAM J. Control and Optimization*, 15(6):959 – 972, 1977.

[25] A. Nekvinda and L. Zajíček. A simple proof of the Rademacker Theorem. *Časopis Pro Pěstování Matematiky*, (4):337 – 341, 1988.

[26] A.S. Nemirovskij and D.B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* Wiley-Interscience, New York, 1983.

[27] A. Nemirovsvki. *Efficient methods in convex programming.* 1994. Lectures notes.

[28] Yu. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, (1):235 – 249, 2005.

[29] Yu. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, (1):127 – 152, 2005.

[30] J. Outrata, M. Kočvara, and J. Zowe. *Nonsmooth Approach to Optimization Problems with Equilibrium Constraints.* Kluwer Academic Publisher, 1998.

[31] B. T. Polyak. *Introduction to optimization.* Optimization Software Inc., 1987.

[32] B.T. Polyak. A general method of solving extremum problems. *Soviet Mathematics Doklady*, 8:593 – 597, 1967.

[33] H. Rademacher. Über partielle und totale Differenzierbarkeit von Funktionen mehrerer Variablen und über die Transformation der Doppelintegrale. *Mat. Ann.*, 89:340–359, 1919.

[34] F. Riesz and B. Sz.-Nagy. *Functional Analysis.* Frederick Ungar Publishing Co., New York, 1955.

[35] R. T. Rockafellar. *Convex analysis.* Princeton University Press, 1970.

[36] H. L. Royden. *Real analysis.* Macmillan Publishing Company, U.S.A., 1988.

[37] H. Schramm and J. Zowe. A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results. *SIAM Journal on Optimization*, 1(2):121–152, 1992.

[38] N.Z. Shor. *Minimization methods for nondifferentiable functions.* Springer-Verlag, Berlin, 1985.

[39] N.Z. Shor, N.G. Zhurbenko, A.P. Likhovid, and P.I. Stetsyuk. Algorithms of nondifferentiable optimization: development and application. *Cybernetics and Systems Analysis*, 39(4):537 – 548, 2003.

[40] K.T. Smith. *Primer of Modern Analysis.* Springer-Verlag, New York, 1971.

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

List of Symbols

$\mathbb{R}$	set of real numbers
$\mathbb{R}_+$	set of non negative real numbers
$\mathbb{R}^n$	set of n-dimensional real vectors
$\mathbb{N}$	set of natural numbers
$\mathbb{Z}$	set of integer numbers
$\mathbb{Q}$	set of rational numbers
(y, z)	open interval: $\{x \in \mathbb{R} : \quad y < x < z\}$
$[y, z]$	closed interval: $\{x \in \mathbb{R} \quad y \leq x \leq z\}$
$\operatorname{aff} C$	affine hull of C
$\operatorname{co} C$	convex hull of C
$\operatorname{cl} C$	closure of C
$\overline{\text{co}}\ C$	closed convex hull of C
$\operatorname{int} C$	interior of C
∂C	boundary of C
$\operatorname{ri} C$	relative interior of C
$\partial_{ri} C$	relative boundary of C
$\operatorname{dist}(x \mid C)$	Euclidean distance from x to C
$\|x\|$	standard Euclidean norm of vector x
$\lvert x \rvert$	absolute value of real number x
$x^T y$	standard inner product
e	vector of all ones in $\mathbb{R}^n$

$B_r^{(n)}(x)$	ball in $\mathbb{R}^n$ of radius r centered at x: $\{y \in \mathbb{R}^n : \quad \|y - x\| \leq r\}$
$B^{(n)} \triangleq B_1^{(n)}(0)$	unit ball in $\mathbb{R}^n$
$S_r^{(n)}(x) \triangleq \partial B_r^{(n)}(x)$	sphere in $\mathbb{R}^n$ of radius r centered at x
$S^{(n)} \triangleq S_1^{(n)}(0)$	unit sphere in $\mathbb{R}^n$
x^T	transpose of vector x
Q^T	transpose of matrix Q
$\lim$	limit
$\limsup$	superior limit
$\liminf$	inferior limit
$f : A \to B$	f is a function on the set A into the set B
$f'_+(x)$	right-hand-side derivative of f at x
$f'_-(x)$	left-hand-side derivative of f at x
$f'(x)$	derivative of f at x
$L_x : \mathbb{R}^n \to \mathbb{R}$	derivative of f at x
$f'(x, d)$	directional derivative of f at x in the direction d
$\nabla f(x)$	gradient of function f at x
$\overline{f}'(x, d)$	upper directional Dini derivative of f at x in the direction d
$\underline{f}'(x, d)$	lower directional Dini derivative of f at x in the direction d
$f^o(x, d)$	Clarke derivative
$\int f$	integral of f
$\operatorname{epi} f$	epigraph of function f
$\operatorname{dom} f$	effective domain of function f
σ_C	support function of C
$\partial f(x)$	subdifferential of f at x
$\partial f(x)$	Clarke gradient of f at x
$\partial_\epsilon^G f(x)$	Goldstein ϵ-subdifferential of f at x
$\operatorname{Conv} \mathbb{R}^n$	set of proper convex functions on $\mathbb{R}^n$ taking values in the extended real axis $\mathbb{R} \cup \{+\infty\}$
$\mathrm{C\overline{onv}}\mathbb{R}^{\mathrm{n}}$	set of closed proper convex functions on $\mathbb{R}^n$ taking values in the extended real axis $\mathbb{R} \cup \{+\infty\}$

C^1	set of continuously differentiable functions on $\mathbb{R}^n$ taking values in the real axis
$\mathrm{BV}[y,z]$	set of functions of bounded variation on $[y,z]$
$\mathcal{L}_n$	Lebesgue measurable sets on $\mathbb{R}^n$
λ_n	Lebesgue measure on $\mathbb{R}^n$
I_n	unit matrix of the order n
$\operatorname*{Argmin}_{x\in\mathcal{C}} f(x)$	set of the minima of f over $\mathcal{C}$
$\operatorname*{argmin}_{x\in\mathcal{C}} f(x)$	minimizer of f over $\mathcal{C}$
$\min_{x\in\mathcal{C}} f(x)$	minimum value of f over $\mathcal{C}$
$\triangleq$	definition operator
$:=$	assignment operator
a.e.	almost everywhere

Ringrazio di cuore i miei tutori, Prof. Manlio Gaudioso e Prof. Maria Flavia Monaco, ed i miei amici Ing. Annabella Astorino e Ing. Antonio Fuduli.

Dedico questa tesi ai miei genitori e ai miei nonni.

Enrico Gorgone