



Selected Topics on Continuous Optimization and Nonsmooth Systems

Jean Charles Gilbert

► **To cite this version:**

Jean Charles Gilbert. Selected Topics on Continuous Optimization and Nonsmooth Systems. Master. Palaiseau, France. 2019, pp.184. cel-01249369v3

HAL Id: cel-01249369

<https://hal.inria.fr/cel-01249369v3>

Submitted on 21 Mar 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Selected Topics on Continuous Optimization and Nonsmooth Systems

Jean Charles GILBERT[†]

Monday 8th March, 2021 (11:00)

[†] INRIA Paris, 2 rue Simone Iff, CS 42112, F-75589 Paris Cedex 12, France; Jean-Charles .
Gilbert@inria.fr.

Preface

This is a preliminary version of the lecture notes of a course given at the University Paris-Saclay in 2015-2021, pompously entitled *Advanced Continuous Optimization*, in *Master classes* simply and adequately named *Optimization*. Despite the use of the qualifier “advanced” in the title of the lectures, their ambition is rather modest, when one looks at the amount of subjects taken up and the level of the research articles, so that we did not keep this qualifier in the title of this document. Naming the course *Intermediate Continuous Optimization* would have been closer to the reality of its contents, but would have attracted less students and less readers! To be more specific, the goal of these lectures is to present and analyze concepts and algorithms in optimization and related domains that are too advanced for being taught in an introductory course, but that should be known by future engineers spending a substantial proportion of their activity in continuous optimization or by students wishing to get involved in the preparation of a PhD thesis in neighboring fields. In a way, these lectures and the present notes can be viewed as an access road to more advanced monographs such as [126, 21, 83, 50, 44, 76], to mention a few of our bookish sources, and a way of filling the gap between basic concepts and research papers in optimization and nonsmooth systems. Rather than being a collection of results, this contribution is meant to be didactic by dedicating a substantial part of its contents to the motivation of concepts and to the presentation of examples and counter-examples. If these lecture notes have awakened the curiosity of the reader or, better still, have deepened his/her knowledge, they will have reached their objective.

This compilation presents both theoretical and numerical subjects, *in a finite dimensional setting*. Whilst many stated results have their counterpart in infinite dimension, we believe that the analysis and algorithmics in finite dimension deserves some interest. First, it is much easier to address than in infinite dimensional spaces, often simpler, and can therefore be taught and learned more rapidly. Furthermore, the knowledge of what is true in finite dimension should allow the reader to make the differences with the infinite dimension corpus and the access to this one should be easier, like linear functional analysis presupposes knowledge in linear algebra. Also, the algorithms, at least those that we shall present in this monograph, find their full usefulness when they are implemented on a computer, which presupposes finite dimension. Readers interested in infinite dimension optimization can pursue the lecture with [21, 23].

The first chapter, entitled *Background*, recalls what should be known to read these lecture notes without difficulty: basic concepts in convex analysis, some less known results in nonsmooth analysis, which is the reason why we devote a little more space on

the generalized differentiability, in theoretical differentiable optimization (optimality conditions and linear optimization duality), and in algorithmics.

Chapter 2, entitled *Optimality Conditions*, starts with more advanced topics and covers the optimality conditions of an optimization problem with constraints expressed by the function inclusion $c(x) \in G$, for some function $c : \mathbb{E} \rightarrow \mathbb{F}$ defined between two Euclidean spaces and a closed convex set $G \subseteq \mathbb{F}$. First order optimality conditions are considered in the first place (section 2.1). There, a large part of the analysis is devoted to Robinson's constraint qualification, which generalizes to the present setting the Mangasarian-Fromovitz constraint qualification for sets defined by equality and inequality constraints. Next, the second order optimality conditions of an optimization problem with equality and inequality constraints is considered (section 2.2). This less often taught matter is important for designing, analyzing, and understanding the behavior of algorithms in nonlinear optimization. They can also be viewed as a preliminary step in the presentation of second order optimality conditions for the more general problem described above, which comes next (section ??).

...

Table of Contents

The chapters and sections marked with the sign “▲” are unfinished, in progress. This does not mean that the unmarked sections are completed; for example, the proof of many results are not provided yet.

This draft of lecture notes is certainly perfectible. Thank you for sending to the author your constructive remarks and suggestions for improvement.

| | | |
|----------|--|----|
| 1 | Background | 9 |
| 1.1 | Notation | 9 |
| 1.2 | Convex Analysis | 10 |
| 1.2.1 | Convex Set | 10 |
| 1.2.2 | Hulls | 11 |
| 1.2.3 | Convex Polyhedron | 13 |
| 1.2.4 | Relative Interior | 13 |
| 1.2.5 | Dual Cone and Farkas Lemma | 15 |
| 1.2.6 | Tangent and Normal Cones | 17 |
| 1.2.7 | Projection | 19 |
| 1.2.8 | Asymptotic Cone | 20 |
| 1.2.9 | Convex Function | 21 |
| 1.2.10 | Asymptotic Function | 23 |
| 1.2.11 | Subdifferentiability | 25 |
| | Notes | 26 |
| | Exercises | 26 |
| 1.3 | Nonsmooth Analysis | 27 |
| 1.3.1 | Lower semi-continuity | 27 |
| 1.3.2 | Lipschitz Continuity | 28 |
| 1.3.3 | Differentiability | 28 |
| 1.3.4 | Multifunction | 29 |
| | Notes | 31 |
| | Exercises | 31 |
| 1.4 | Optimization | 31 |
| 1.4.1 | Generic Problem | 31 |
| 1.4.2 | Peano-Kantorovich Optimality Condition | 32 |
| 1.4.3 | Equality Constrained Problem (P_E) | 34 |
| 1.4.4 | Equality and Inequality Constrained Problem (P_{EI}) | 37 |
| 1.4.5 | Abstract Duality | 41 |
| 1.4.6 | Linear Optimization Problem (P_L) | 43 |

| | | |
|----------|---|-----------|
| 1.5 | Algorithmics | 45 |
| 1.5.1 | Speeds of Convergence | 45 |
| 1.5.2 | Newton and Quasi-Newton Algorithms ▲ | 46 |
| 1.5.3 | Global Convergence in Unconstrained Optimization ▲ | 47 |
| 1.5.4 | Global Convergence for Nonlinear Equations ▲ | 49 |
| 2 | Optimality Conditions | 51 |
| 2.1 | First Order Optimality Conditions for (P_G) | 51 |
| 2.1.1 | Definition of the General Problem | 51 |
| 2.1.2 | First Order Optimality Conditions | 53 |
| 2.1.3 | Robinson's Condition | 57 |
| 2.1.4 | Robinson's Constraint Qualification | 71 |
| 2.1.5 | Set of optimal multipliers | 73 |
| 2.1.6 | Other problems ▲ | 76 |
| | Notes | 79 |
| | Exercises | 80 |
| 2.2 | Second Order Optimality Conditions for (PEI) | 82 |
| 2.2.1 | Critical Cone | 83 |
| 2.2.2 | Three Instructive Examples | 85 |
| 2.2.3 | Second Order Necessary Optimality Conditions | 86 |
| 2.2.4 | Second Order Sufficient Optimality Conditions | 90 |
| | Notes | 92 |
| | Exercises | 92 |
| 3 | Perturbation Analysis | 95 |
| 3.1 | Linear System ▲ | 95 |
| 3.2 | Nonlinear System ▲ | 95 |
| 3.3 | Optimization ▲ | 95 |
| | Notes | 96 |
| 4 | A Few Methods for Nonsmooth Systems | 97 |
| 4.1 | Josephy-Newton Algorithm for Functional Inclusions | 98 |
| 4.1.1 | A Gallery of Problems | 98 |
| 4.1.2 | The Josephy-Newton Algorithm | 101 |
| 4.1.3 | Regularity | 102 |
| 4.1.4 | Speed of Convergence | 108 |
| 4.1.5 | Local Convergence | 110 |
| 4.1.6 | Globalization by Line-Search ▲ | 112 |
| | Notes | 112 |
| | Exercises | 112 |
| 4.2 | B-Newton Method for Systems of Equations ▲ | 114 |
| 4.3 | Linearization method for \mathcal{PC}^1 functions ▲ | 114 |
| 4.4 | Semismooth Newton Method for nonlinear systems | 114 |
| 4.4.1 | Motivation, Orientation, Examples | 114 |
| 4.4.2 | Generalized Differentiability | 118 |
| 4.4.3 | Semismoothness ▲ | 122 |
| 4.4.4 | The Semismooth Newton Method ▲ | 126 |

| | | |
|-------------------|--|-----|
| 4.4.5 | Globalization by linesearch ▲ | 127 |
| 4.4.6 | Globalization by trust regions ▲ | 128 |
| 4.4.7 | Examples of use ▲ | 128 |
| | Notes | 128 |
| | Exercises | 128 |
| 4.5 | Reformulation Methods for Complementarity Problems ▲ | 128 |
| 4.5.1 | Fischer-Newton Algorithm | 130 |
| 4.5.2 | Newton-Min Algorithm | 131 |
| | Notes | 131 |
| 5 | A Few Methods for Optimization | 133 |
| 5.1 | SQP Algorithm for (P_{EI}) | 133 |
| 5.1.1 | The SQP Algorithm | 134 |
| 5.1.2 | Local Convergence | 138 |
| 5.1.3 | Exact Penalization | 141 |
| 5.1.4 | Globalization by Line-Search for (P_{EI}) ▲ | 146 |
| 5.1.5 | Globalization by Trust-Region for (P_E) ▲ | 152 |
| | Notes ▲ | 152 |
| | Exercises | 152 |
| 5.2 | SQP Algorithm for (P_G) ▲ | 153 |
| 5.2.1 | The SQP Algorithm | 153 |
| 5.2.2 | Local Convergence | 154 |
| 6 | Self-dual Conic Optimization | 155 |
| 6.1 | Semidefinite Optimization | 156 |
| 6.1.1 | Primal and Dual Problems | 156 |
| 6.1.2 | Examples of SDO Formulation | 160 |
| 6.1.3 | Existence of Solution, Optimality Conditions | 163 |
| 6.1.4 | Interior Point Algorithms ▲ | 167 |
| 6.1.5 | A Nonsmooth Algorithm ▲ | 170 |
| | Notes | 170 |
| | Exercises | 170 |
| 6.2 | Circular Optimization ▲ | 170 |
| 6.3 | Copositive Optimization ▲ | 170 |
| <hr/> | | |
| Appendices | | |
| <hr/> | | |
| A | References | 173 |
| B | Index | 181 |

1 Background

In these notes, vector spaces are always assumed to have finite dimension; they are usually denoted by the letters $\mathbb{E}, \mathbb{F}, \mathbb{G}, \dots$

1.1 Notation

The set of nonnegative integers is denoted by \mathbb{N} and an interval of integers is denoted by

$$[n_1 : n_2] := \{n_1, \dots, n_2\},$$

where we have assumed that the integers n_1 and n_2 verify $n_1 \leq n_2$.

The set of real numbers is denoted by \mathbb{R} and we note

$$\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, +\infty\}, \quad \mathbb{R}_+ := \{t \in \mathbb{R} : t \geq 0\}, \quad \mathbb{R}_{++} := \{t \in \mathbb{R} : t > 0\},$$

$\mathbb{R}_- := -\mathbb{R}_+$, and $\mathbb{R}_{--} := -\mathbb{R}_{++}$. For $a \leq b$ in \mathbb{R} , one defines the *intervals*

$$\begin{aligned} [a, b] &:= \{t \in \mathbb{R} : a \leq t \leq b\}, \\ [a, b) &:= \{t \in \mathbb{R} : a \leq t < b\}, \\ (a, b] &:= \{t \in \mathbb{R} : a < t \leq b\}, \\ (a, b) &:= \{t \in \mathbb{R} : a < t < b\}. \end{aligned}$$

Hence, the last three intervals are empty if $a = b$.

The set of real vectors of integer dimension $n \geq 0$ is denoted by \mathbb{R}^n . Inequalities in \mathbb{R}^n must be understood componentwise; hence, for a and $b \in \mathbb{R}^n$, $a \leq b$ (resp. $a < b$) means $a_i \leq b_i$ (resp. $a_i < b_i$) for all $i \in [1 : n]$. We note

$$\mathbb{R}_+^n := \{x \in \mathbb{R}^n : x \geq 0\}, \quad \mathbb{R}_{++}^n := \{x \in \mathbb{R}^n : x > 0\},$$

$\mathbb{R}_-^n := -\mathbb{R}_+^n$, and $\mathbb{R}_{--}^n := -\mathbb{R}_{++}^n$.

The *Minkowski sum* of two subsets P and Q of a vector space \mathbb{E} is denoted and defined by

$$P + Q := \{x + y : x \in P, y \in Q\}.$$

When P is the singleton $\{x\}$, we simply write $x + Q$ for $\{x\} + Q$. For $\Lambda \subseteq \mathbb{R}$ and $P \subseteq \mathbb{E}$, we note

$$\Lambda P := \{\lambda x : \lambda \in \Lambda, x \in P\}.$$

When Λ is the singleton $\{\lambda\}$, we simply note λP for $\{\lambda\}P$. Note that $P - P$ contains 0 but is usually not reduced to $\{0\}$, unless P is a singleton. Similarly, $P + P$ contains $2P$,

but the two sets are usually different, unless P is convex (definition in section 1.2.1 below).

If \mathbb{E} and \mathbb{F} are two vector spaces, the *adjoint* of a linear map $A : \mathbb{E} \rightarrow \mathbb{F}$ is the linear map $A^* : \mathbb{F} \rightarrow \mathbb{E}$ uniquely defined by:

$$\forall (x, y) \in \mathbb{E} \times \mathbb{F} : \quad \langle A^*y, x \rangle = \langle y, Ax \rangle.$$

Because of their finite dimension, there is no restriction in assuming that vector spaces are equipped with a norm, denoted by $\|\cdot\|$, or even with a scalar product, denoted by $\langle \cdot, \cdot \rangle$. It is the topology associated with this norm or scalar product that is assumed to equip the vector space. The open and closed unit balls centered at the origin are then denoted by

$$B := \{x \in \mathbb{E} : \|x\| < 1\} \quad \text{and} \quad \bar{B} := \{x \in \mathbb{E} : \|x\| \leq 1\}.$$

For $x \in \mathbb{E}$ and $r > 0$, we also use the notation $B(x, r) := x + rB$ and $\bar{B}(x, r) := x + r\bar{B}$ for the open and closed balls of radius r , centered at x . The *interior* of a set $S \subseteq \mathbb{E}$ is denoted by $\text{int}(S)$, $\text{int } S$, or S° , and its *closure* by $\text{cl}(S)$, $\text{cl } S$, or \bar{S} . The set of *neighborhoods* of a point $x \in \mathbb{E}$ is denoted by $\mathcal{N}(x)$.

In a normed space $(\mathbb{E}, \|\cdot\|)$, the *distance to a set*, say $S \subseteq \mathbb{E}$, is denoted and defined at $x \in \mathbb{E}$ by

$$\text{dist}(x, S) := \inf_{x' \in S} \|x' - x\|. \quad (1.1)$$

By definition of the infimum, this distance is infinite when $S = \emptyset$.

When \mathbb{E} is a Euclidean vector space, the *gradient* of a function $f : \mathbb{E} \rightarrow \mathbb{R}$ at x , denoted $\nabla f(x)$, is the unique vector of \mathbb{E} defined from the derivative $f'(x)$ by

$$\langle \nabla f(x), d \rangle = f'(x) \cdot d, \quad \forall d \in \mathbb{E}.$$

Note that the gradient depends on the chosen scalar product of \mathbb{E} , which is not the case for the derivative.

1.2 Convex Analysis

1.2.1 Convex Set

Let \mathbb{E} be a finite dimensional vector space. With two points x_0 and x_1 of \mathbb{E} , one can form the following *segments* of \mathbb{E} :

$$\begin{aligned} [x_0, x_1] &:= \{(1-t)x_0 + tx_1 : t \in [0, 1]\}, \\]x_0, x_1[&:= \{(1-t)x_0 + tx_1 : t \in [0, 1)\}, \\ (x_0, x_1] &:= \{(1-t)x_0 + tx_1 : t \in (0, 1]\}, \\]x_0, x_1) &:= \{(1-t)x_0 + tx_1 : t \in (0, 1)\}. \end{aligned}$$

Hence, when $x_0 = x_1$, these four segments are reduced to the single point $\{x_0\}$.

A set $C \subseteq \mathbb{E}$ is *convex* if

$$\forall (x_0, x_1) \in C^2 : \quad [x_0, x_1] \subseteq C.$$

Convex Sets Calculus

Here are some other convex sets encountered in these notes.

- The **Minkowski sum** $C_1 + C_2$ of two convex sets C_1 and C_2 is a convex set.
- The product $\alpha C := \{\alpha x : x \in C\}$ of a scalar $\alpha \in \mathbb{R}$ by a convex set C is a convex set.
- If $\{C_i\}_{i \in I}$ is an arbitrary family of convex sets of a vector space \mathbb{E} , then their intersection $\bigcap_{i \in I} C_i$ is a convex set (but not their union!).
- If $A : \mathbb{E} \rightarrow \mathbb{F}$ is a linear map between two vector spaces \mathbb{E} and \mathbb{F} , the direct image $A(C) := \{Ax : x \in C\}$ (resp. the inverse image $A^{-1}(C) := \{x \in \mathbb{E} : Ax \in C\}$) of a convex set C of \mathbb{E} (resp. of \mathbb{F}) by A is a convex set.
- Let $C_1 \subseteq \mathbb{E}_1$ and $C_2 \subseteq \mathbb{E}_2$. Then, $C_1 \times C_2$ is convex in $\mathbb{E}_1 \times \mathbb{E}_2$ if and only if C_1 and C_2 are convex.

Examples of Convex Sets

Here are some other convex sets encountered in these notes.

- The *unit simplex* of \mathbb{R}^n is the set

$$\Delta_n := \{x \in \mathbb{R}^n : e^\top x = 1, x \geq 0\},$$

where $e = (1, \dots, 1) \in \mathbb{R}^n$.

- The **cone** (see below) of positive semidefinite symmetric matrices of order n is convex and denoted by \mathcal{S}_+^n . The notation $M \geq 0$ means that $M \in \mathcal{S}_+^n$.
- The **cone** of positive definite symmetric matrices of order n is convex and denoted by \mathcal{S}_{++}^n . The notation $M > 0$ means that $M \in \mathcal{S}_{++}^n$.

1.2.2 Hulls

Affine and Vector Hulls

The *affine hull* of an arbitrary set $P \subseteq \mathbb{E}$ is the smallest affine space containing P :

$$\text{aff } P := \bigcap \{A : A \text{ is an affine space containing } P\}.$$

This definition makes sense since the intersection of an arbitrary collection of affine spaces of \mathbb{E} is an affine space of \mathbb{E} . It can be shown that

$$\text{aff } P = \left\{ \sum_{i=1}^m \alpha_i x_i : m \in \mathbb{N}, \text{ all } x_i \in P, \alpha \in \mathbb{R}^m, e^\top \alpha = 1 \right\},$$

where we have denoted by e the vector of all ones.

The *vector hull* of an arbitrary set $P \subseteq \mathbb{E}$ is the smallest subspace of \mathbb{E} containing P :

$$\text{vect } P := \bigcap \{E : E \text{ is a subspace containing } P\}.$$

This definition makes sense since the intersection of an arbitrary collection of subspaces of \mathbb{E} is a subspace of \mathbb{E} . It is not difficult to see that

$$\text{vect } P = \text{aff}(P \cup \{0\}) = \left\{ \sum_{i=1}^m \alpha_i x_i : m \in \mathbb{N}, \text{ all } x_i \in P, \alpha \in \mathbb{R}^m \right\}.$$

Convex Hull

The *convex hull* of an arbitrary set $P \subseteq \mathbb{E}$ is the smallest convex set containing P :

$$\text{co } P := \bigcap \{C : C \text{ is a convex set containing } P\}.$$

This definition makes sense since the intersection of an arbitrary collection of convex sets of \mathbb{E} is a convex set of \mathbb{E} . It can be shown that

$$\text{co } P = \left\{ \sum_{i=1}^m \alpha_i x_i : m \in \mathbb{N}, \text{ all } x_i \in P, \alpha \in \mathbb{R}_+^m, e^\top \alpha = 1 \right\}.$$

In finite dimension, one can limit m in the previous sum to $\dim \mathbb{E} + 1$ (Carathéodory [29; 1907]) and this observation yields the following important implication:

$$\boxed{P \text{ is compact}} \implies \boxed{\text{co } P \text{ is compact.}} \quad (1.2)$$

But, even if P is closed, $\text{co } P$ may not be closed, which motivates the introduction of the following concept.

Closed Convex Hull

The *closed convex hull* of an arbitrary set $P \subseteq \mathbb{E}$ is the smallest closed convex set containing P :

$$\overline{\text{co}} P := \bigcap \{C : C \text{ is a closed convex set containing } P\}.$$

This definition makes sense since the intersection of an arbitrary collection of closed convex sets of \mathbb{E} is a closed convex set of \mathbb{E} . It can be shown that

$$\boxed{\overline{\text{co}} P = \overline{\text{co } P}.$$

Conic Hull

Recall that a part K of a vector space \mathbb{E} is a *cone* if $\mathbb{R}_{++} K \subseteq K$, which means that there must hold $tx \in K$ each time $t > 0$ and $x \in K$. Hence a cone may or may not contain zero (this is the reason why t is taken in \mathbb{R}_{++} above and not in \mathbb{R}_+), which allows us to talk about the cone of positive definite symmetric matrices. The *conical hull* of an arbitrary set $P \subseteq \mathbb{E}$ is the smallest *convex* cone containing P ¹:

$$\text{cone } P := \bigcap \{K : K \text{ is a convex cone containing } P\}.$$

This definition makes sense since the intersection of an arbitrary collection of convex cones of \mathbb{E} is a convex cone of \mathbb{E} . It can be shown that

$$\text{cone } P = \left\{ \sum_{i=1}^m \alpha_i x_i : m \in \mathbb{N}, \text{ all } x_i \in P, \alpha \in \mathbb{R}_+^m \right\}.$$

¹ Rockafellar [125; p. 14] finds it convenient to add the origin to $\text{cone } P$, which makes no difference with our definition when $0 \in P$.

1.2.3 Convex Polyhedron

A *convex polyhedron* of a vector space \mathbb{E} is a set of the form

$$P = \{x \in \mathbb{E} : Ax \leq b\}, \quad (1.3)$$

where $A : \mathbb{E} \rightarrow \mathbb{R}^m$ is a linear map and $b \in \mathbb{R}^m$. A convex set that can be written as the set P above is said to be *polyhedral*. A convex polyhedron is therefore an intersection of a finite number of half spaces, namely $\{x \in \mathbb{E} : (Ax - b)_i \leq 0\}$ for $i \in [1 : m]$, hence it is closed and convex. For P of the form (1.3) and for $x \in P$, one defines

$$I(x) := \{i \in [1 : m] : (Ax - b)_i = 0\}. \quad (1.4)$$

It is clear that

$$(\{x_k\} \rightarrow x) \implies I(x_k) \subseteq I(x) \text{ for } k \text{ large.} \quad (1.5)$$

The representation of a polyhedron by (1.3) is qualified as *dual*, since it involves linear operators (hence elements of the dual space of \mathbb{E}). Such a set has also a *primal representation*, which makes use of convex and conic hulls. Indeed, the set P in (1.3) can also be written in the following form

$$P = \text{co}\{x_1, \dots, x_p\} + \text{cone}\{y_1, \dots, y_q\}, \quad (1.6)$$

where $\{x_i\}_{i \in [1 : p]}$ are points of \mathbb{E} and $\{y_j\}_{j \in [1 : q]}$ are “directions” of \mathbb{E} (another name for a point of \mathbb{E} , to quote that it is used here to generate a cone). The converse is also true: any set of the form (1.6) can be written in the form (1.3). This equivalence between these representations of a convex polyhedron has been shown by Minkowski.

Proposition 1.1 (polyhedrality properties)

- 1) Linear transformation *if* $P \subseteq \mathbb{E}$ *is a convex polyhedron and* $T : \mathbb{E} \rightarrow \mathbb{F}$ *is linear, then* $T(P) \subseteq \mathbb{F}$ *is a convex polyhedron.*
- 2) Addition: *if* P_1 *and* P_2 *are polyhedra, then* $P_1 + P_2$ *is a polyhedron.*
- 3) Upper semi-continuity of I : *for* $x \in P$, $\exists \delta > 0$ *such that* $x' \in B(x, \delta) \cap P$ *implies that* $I(x') \subseteq I(x)$.

1.2.4 Relative Interior

The *relative interior* of $P \subseteq \mathbb{E}$ is its interior in $\text{aff } P$ equipped with the relative topology, the one induced from that of \mathbb{E} :

$$\text{ri } P := \{x \in P : \exists r > 0 \text{ such that } [B(x, r) \cap \text{aff } P] \subseteq P\}.$$

Note that $P_1 \subseteq P_2$ does not necessarily imply that $\text{ri } P_1 \subseteq \text{ri } P_2$, but one has

$$P_1 \subseteq P_2 \quad \text{and} \quad \text{aff } P_1 = \text{aff } P_2 \quad \implies \quad \text{ri } P_1 \subseteq \text{ri } P_2. \quad (1.7)$$

In *finite dimension*, the following holds

$$C \text{ convex and nonempty} \implies \begin{cases} \text{ri } C \neq \emptyset, \\ \text{aff}(\text{ri } C) = \text{aff } C. \end{cases} \quad (1.8)$$

Lemma 1.2 *Let C be a nonempty convex set and $x \in \mathbb{E}$. Then,*

$$x \in \text{ri } C \text{ and } y \in \overline{C} \implies [x, y] \subseteq \text{ri } C.$$

Proposition 1.3 (relative interior criterion) *Let C be a nonempty convex set and $x \in \mathbb{E}$. Then,*

$$x \in \text{ri } C \iff \forall x_0 \in C, \exists t > 1 : (1-t)x_0 + tx \in C, \quad (1.9a)$$

$$\iff \forall x_0 \in \text{aff } C, \exists t > 1 : (1-t)x_0 + tx \in C. \quad (1.9b)$$

Let C be a nonempty convex set. Then,

$$\text{ri } C \text{ and } \overline{C} \text{ are convex,}$$

$$\text{aff } \overline{C} = \text{aff } C,$$

$$\overline{\text{ri } C} = \overline{C} \quad \text{and} \quad \text{ri } \overline{C} = \text{ri } C. \quad (1.10)$$

In short, the last identities tell us that, when the two operators “ri” and “cl” act in sequence on C , it is always the last one acting that prevails.

Proposition 1.4 (relative interior calculus) *Let $\mathbb{E}, \mathbb{E}_1, \mathbb{E}_2$, and \mathbb{F} be vector spaces and $A : \mathbb{E} \rightarrow \mathbb{F}$ be a linear map.*

1) (Cartesian product) *If $C_1 \subseteq \mathbb{E}_1$ and $C_2 \subseteq \mathbb{E}_2$ are convex sets, then*

$$\text{ri}(C_1 \times C_2) = (\text{ri } C_1) \times (\text{ri } C_2).$$

2) (Intersection) *If $(C_i)_{i \in I}$ is a family of convex sets in \mathbb{E} such that $\cap_{i \in I} \text{ri } C_i$ is nonempty, then*

$$\text{ri}(\cap_{i \in I} C_i) \subseteq \cap_{i \in I} \text{ri } C_i. \quad (1.11)$$

with equality if I is finite.

3) (Linear map) *If $C \subseteq \mathbb{E}$ is a convex set, then*

$$\text{ri}(A(C)) = A(\text{ri } C). \quad (1.12)$$

4) (Linear preimage) If $C \subseteq \mathbb{F}$ is a convex set and if the preimage $A^{-1}(\text{ri } C) \neq \emptyset$, then

$$\text{ri}(A^{-1}(C)) = A^{-1}(\text{ri } C).$$

5) (Multiplication) If $C \subseteq \mathbb{E}$ is a convex set and $\alpha \in \mathbb{R}$, then

$$\text{ri}(\alpha C) = \alpha (\text{ri } C).$$

6) (Sum) If $C_1 \subseteq \mathbb{E}$ and $C_2 \subseteq \mathbb{E}$ are convex sets, then

$$\text{ri}(C_1 + C_2) = (\text{ri } C_1) + (\text{ri } C_2). \quad (1.13)$$

A point x in a subset $P \subseteq \mathbb{E}$ is said to be *absorbing*² for P if $\forall d \in \mathbb{E}, \exists t > 0$ such that $x + td \in P$. In *finite dimension* and for a convex set C , one can certify the interiority of a point by looking along all the directions, since the following holds

$$x \in \text{int } C \iff x \text{ is absorbing for } C. \quad (1.14)$$

1.2.5 Dual Cone and Farkas Lemma

Let P be a subset of a Euclidean vector space \mathbb{E} . The (*positive*) *dual cone* of P is defined and denoted by

$$P^+ := \{d \in \mathbb{E} : \langle d, x \rangle \geq 0, \forall x \in P\}.$$

It is a nonempty close convex cone. When $P \equiv \mathbb{E}_0$ is a subspace of \mathbb{E} , \mathbb{E}_0^+ is the subspace orthogonal to \mathbb{E}_0 , denoted by $\mathbb{E}_0^\perp := \{d \in \mathbb{E} : \langle d, x \rangle = 0, \forall x \in \mathbb{E}_0\}$. A cone is said to be *self-dual* if $K^+ = K$. The *negative dual cone* of a set P is

$$P^- := -P^+.$$

The *bidual cone* of P , denoted P^{++} , is the dual cone of the dual cone of P :

$$P^{++} := (P^+)^+.$$

Points in the interior (if any) and relative interior of P^+ are characterized in exercise 1.2.10.

The next lemma is of paramount importance for chapter 2. It gives a description of the closure of the linear image of a convex cone, using dual cones. We have denoted by A^* the [adjoint](#) of the linear map A .

² The terms *absorbent point* are also used instead of *absorbing point*.

Lemma 1.5 (Farkas, generalized) *Let \mathbb{E} and \mathbb{F} be two Euclidean spaces, $A : \mathbb{E} \rightarrow \mathbb{F}$ be a linear map, and K be a nonempty convex cone of \mathbb{E} . Then,*

$$\overline{A(K)} = \{y \in \mathbb{F} : A^*y \in K^+\}^+. \quad (1.15)$$

Remarks 1.6 1) In general, one cannot get rid of the closure on $A(K)$ in the left-hand side of (1.15), since the dual cone in the right-hand side is closed, while the linear map of a nonempty (even closed) convex cone is not necessarily closed. Here is a counter-example, using the *circular cone* $K = \mathbb{R}_{\nabla}^3$ (see also section 6.2): the image of

$$\mathbb{R}_{\nabla}^3 = \{x \in \mathbb{R}^3 : x_3 \geq \|(x_1, x_2)\|_2\} \quad \text{by} \quad A : \mathbb{R}^3 \rightarrow \mathbb{R}^2 : x \mapsto (x_1, x_2 + x_3)$$

is the non closed cone $A(\mathbb{R}_{\nabla}^3) = \{x \in \mathbb{R}^2 : x_2 > 0\} \cup \{(0, 0)\}$.

- 2) If K is polyhedral, then $A(K)$ is polyhedral (a property recalled in section 1.2.3), hence closed, so that the closure in the left-hand side of (1.15) can be discarded (see exercise 1.2.8).
- 3) For $K = \mathbb{E}$, one recovers the linear algebra identity $\mathcal{R}(A) = \mathcal{N}(A^*)^\perp$.
- 4) The importance of the Farkas lemma lies also on the fact that it is an *existence* result. This is more easily described when $A(K)$ is closed. Then, the identity (1.15) tells us that, when a vector d has a certain property (i.e., it has a nonnegative scalar product with any vector y such that $A^*y \in K^+$), there *exists* a vector $x \in K$ such that $d = Ax$. In optimization, one uses this lemma to prove the existence of optimal multipliers (see the proof of theorems 1.40 and 2.6).
- 5) Assume that K is a nonempty convex cone. Applying the Farkas identity (1.15) with $A = I$, the identity in $\mathbb{E} = \mathbb{F}$, yields

$$K^{++} = \overline{K}. \quad (1.16)$$

- 6) The proof of the Farkas identity (1.15) can be obtained by a separation argument (if the inclusion \supseteq does not hold, one separates $\overline{A(K)}$ and a point not belonging to it by a proper hyperplane). Now, it is not difficult to obtain (1.15) as a consequence of (1.16), which has been viewed as a consequence of (1.15)! Actually, (1.16) could also be proved by a separation argument. To deduce (1.15) from (1.16), first observe that

$$A(K)^+ = \{y \in \mathbb{F} : A^*y \in K^+\}, \quad (1.17)$$

which is easy to prove. Then, take the dual of both sides of this identity and observe that, since $A(K)$ is a convex cone, its bidual is its closure by (1.16). The identity (1.15) follows. \square

Proposition 1.7 (dual cone calculus) *Let \mathbb{E} be a Euclidean vector space.*

1) *If P_1 and P_2 are two nonempty subsets of \mathbb{E} , then*

$$P_1 \subseteq P_2 \quad \Longrightarrow \quad P_1^+ \supseteq P_2^+.$$

2) *If P is a nonempty subset of \mathbb{E} , then*

$$P^+ = (\mathbb{R}_+ P)^+ = (\text{co } P)^+ = (\text{cl } P)^+. \quad (1.18)$$

3) *If P is a nonempty subset of \mathbb{E} , then*

$$P^{++} = \overline{\text{co}}(\mathbb{R}_+ P). \quad (1.19)$$

In particular,

$$P \subseteq P^{++},$$

with equality if and only if P is a nonempty closed convex cone.

4) *For nonempty subsets P_1 and P_2 of \mathbb{E} , one has*

$$(P_1 + P_2)^+ \supseteq P_1^+ \cap P_2^+,$$

with equality if $0 \in \overline{P_1} \cap \overline{P_2}$. In particular, if K_1 and K_2 are two nonempty cones, one has

$$(K_1 + K_2)^+ = K_1^+ \cap K_2^+. \quad (1.20)$$

5) *If K_1 and K_2 are two nonempty closed convex cones, then*

$$(K_1 \cap K_2)^+ = \overline{K_1^+ + K_2^+}. \quad (1.21)$$

6) *If $(P_i)_{i \in I}$ is an arbitrary family of nonempty subsets P_i of \mathbb{E} , then*

$$\left(\bigcup_{i \in I} P_i \right)^+ = \bigcap_{i \in I} P_i^+.$$

7) *If $(\mathbb{E}_1, \langle \cdot, \cdot \rangle_1)$ and $(\mathbb{E}_2, \langle \cdot, \cdot \rangle_2)$ are two Euclidean spaces, if $\mathbb{E}_1 \times \mathbb{E}_2$ is equipped with the scalar product $\langle (x_1, x_2), (y_1, y_2) \rangle = \langle x_1, y_1 \rangle_1 + \langle x_2, y_2 \rangle_2$, and if $\emptyset \neq Q_1 \subseteq \mathbb{E}_1$ and $\emptyset \neq Q_2 \subseteq \mathbb{E}_2$, then*

$$(Q_1 \times Q_2)^+ \supseteq Q_1^+ \times Q_2^+,$$

with equality when $0 \in \text{cl}(Q_1) \cap \text{cl}(Q_2)$.

1.2.6 Tangent and Normal Cones

Let C be a closed convex set of a vector space \mathbb{E} . The *cone of feasible directions* to C at $x \in C$ is defined and denoted by

$$\mathbf{T}_x^f C \equiv \mathbf{T}_C^f(x) := \mathbb{R}_+(C - x).$$

An element of that cone is called a *feasible direction*; such a direction $d \in \mathbb{E}$ is therefore characterized by the fact that there is a $t > 0$ such that $x + td \in C$ (or, by the convexity of C , $x + td \in C$ for all sufficiently small $t > 0$). Note that, by convexity of C , $t(C - x) \subseteq C - x$ when $t \in [0, 1]$, so that one also have $\mathbf{T}_C^f(x) := [1, \infty)(C - x)$.

The cone of feasible directions is usually not a closed set. For example, when C is the unit closed ball of \mathbb{R}^2 , $\mathbf{T}_{(0,-1)} C = \{x \in \mathbb{R}^2 : x_2 > 0\} \cup \{(0, 0)\}$, which is not closed. The *tangent cone* to C at $x \in C$ is the closure of the cone of feasible directions. It is denoted by

$$\mathbf{T}_x C \equiv \mathbf{T}_C(x) = \overline{\mathbb{R}_+(C - x)}. \quad (1.22)$$

When $x \notin C$, one sets $\mathbf{T}_C^f(x) = \emptyset$ and $\mathbf{T}_C(x) = \emptyset$. The cone of feasible directions is not necessary the relative interior of the tangent cone. For example, as indicated below, when C is a convex polyhedron, $\mathbf{T}_x^f C$ is identical to the tangent cone $\mathbf{T}_x C$, which is also a convex polyhedron, hence (relatively) closed.

Now, let C be a closed convex set of a *Euclidean* vector space \mathbb{E} . The *normal cone* to C at $x \in C$ is defined and denoted by

$$\mathbf{N}_x C \equiv \mathbf{N}_C(x) = \{d \in \mathbb{E} : \langle x' - x, d \rangle \leq 0, \forall x' \in C\}. \quad (1.23)$$

We also set $\mathbf{N}_C(x) = \emptyset$ when $x \notin C$. The following hold

$$\mathbf{N}_x C = (\mathbf{T}_x C)^- \quad \text{and} \quad \mathbf{T}_x C = (\mathbf{N}_x C)^-.$$

Proposition 1.8 (tangent and normal cone calculus)

1) (Intersection) *If C_1 and C_2 are closed convex sets of a vector space \mathbb{E} and $x \in C_1 \cap C_2$, then*

$$\mathbf{T}_x(C_1 \cap C_2) \subseteq \mathbf{T}_x C_1 \cap \mathbf{T}_x C_2, \quad (1.24a)$$

$$\mathbf{N}_x(C_1 \cap C_2) \supseteq \mathbf{N}_x C_1 + \mathbf{N}_x C_2, \quad (1.24b)$$

with equalities if $0 \in \text{ri}(C_1 - C_2)$ or $(\text{ri} C_1) \cap (\text{ri} C_2) \neq \emptyset$.

2) (Product) *Let $(\mathbb{E}_1, \langle \cdot, \cdot \rangle_1)$ and $(\mathbb{E}_2, \langle \cdot, \cdot \rangle_2)$ be two Euclidean spaces and equip $\mathbb{E}_1 \times \mathbb{E}_2$ with the scalar product $\langle (x_1, x_2), (y_1, y_2) \rangle = \langle x_1, y_1 \rangle_1 + \langle x_2, y_2 \rangle_2$. If C_1 (resp. C_2) is a closed convex set of \mathbb{E}_1 (resp. \mathbb{E}_2), then $C_1 \times C_2 := \{(x_1, x_2) : x_1 \in C_1, x_2 \in C_2\}$ is a closed convex set of $\mathbb{E}_1 \times \mathbb{E}_2$ and at $(x_1, x_2) \in C_1 \times C_2$, one has*

$$\mathbf{T}_{(x_1, x_2)}(C_1 \times C_2) = (\mathbf{T}_{x_1} C_1) \times (\mathbf{T}_{x_2} C_2), \quad (1.25a)$$

$$\mathbf{N}_{(x_1, x_2)}(C_1 \times C_2) = (\mathbf{N}_{x_1} C_1) \times (\mathbf{N}_{x_2} C_2). \quad (1.25b)$$

The tangent and normal cones to the convex polyhedron $P = \{x \in \mathbb{E} : Ax \leq b\}$ has the following expressions and properties. Recall the notation (1.4): $I(x) := \{i \in [1 : m] : (Ax - b)_i = 0\}$.

- *Tangent cones:*

$$\mathbb{T}_P(x) = \mathbb{T}_P^f(x) = \{d \in \mathbb{E} : (Ad)_{I(x)} \leq 0\}, \quad (1.26a)$$

$$I(x_1) \subseteq I(x_2) \implies \mathbb{T}_P(x_1) \supseteq \mathbb{T}_P(x_2). \quad (1.26b)$$

- *Normal cone:*

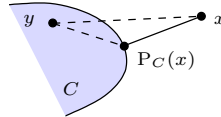
$$\mathbb{N}_P(x) = \text{cone}\{A^*e_i : i \in I(x)\}, \quad (1.27a)$$

$$I(x_1) \subseteq I(x_2) \implies \mathbb{N}_P(x_1) \subseteq \mathbb{N}_P(x_2). \quad (1.27b)$$

We see from (1.26a) and (1.27a) that the tangent and normal cones to a convex polyhedron are polyhedral.

1.2.7 Projection

Let \mathbb{E} be a *Euclidean* space, with a scalar product denoted by $\langle \cdot, \cdot \rangle$ and its associated norm denoted by $\| \cdot \|$. For a *nonempty, closed, convex* set C in \mathbb{E} , the problem



$$\inf_{y \in C} \|y - x\| \quad (1.28)$$

has a *unique* solution, called the (*orthogonal*) *projection* of x on C . This projection is denoted by $P_C(x)$. The function $P_C : \mathbb{E} \rightarrow C$ is called the (*orthogonal*) *projector* on C .

Proposition 1.9 (characterizations of the projection) For $x \in \mathbb{E}$ and $\bar{x} \in C$, one has

$$\bar{x} = P_C(x) \iff \langle y - \bar{x}, \bar{x} - x \rangle \geq 0, \quad \forall y \in C \quad (1.29a)$$

$$\iff \langle y - \bar{x}, y - x \rangle \geq 0, \quad \forall y \in C \quad (1.29b)$$

$$\iff \langle y - x, \bar{x} - x \rangle \geq \|\bar{x} - x\|^2, \quad \forall y \in C. \quad (1.29c)$$

Remarks 1.10 1) The scalar products in the right-hand sides of the equivalences (1.29) use each time two vectors among the possible three $\bar{x} - x$, $y - \bar{x}$, and $y - x$. To have a characterization of the projection, it suffices to have a nonnegative scalar product of $y - \bar{x}$ with one of the two other vectors, like in (1.29a) and (1.29b). When $y - \bar{x}$ is not present, the inequality must be strengthened, like in (1.29c) to have a characterization, as shown by the counter-example where $C = [0, 1] \subseteq \mathbb{R}$, $x \notin C$, while y and \bar{x} are chosen arbitrarily in C (hence \bar{x} is not necessarily the projection of x).

2) The most often used characterization is (1.29a). Maybe that this is because it also expresses the optimality condition “ $f'(\bar{x}; y - \bar{x}) \geq 0$, for all $y \in C$ ” of the optimization problem in (1.28), defining the projection, rewritten $\inf\{f(y) := \frac{1}{2}\|y - x\|^2 : y \in C\}$. □

Proposition 1.11 (properties of the projection) *The projection on a nonempty closed convex set C has the following properties:*

- 1) $\forall (x_1, x_2) \in \mathbb{E}^2, \langle P_C(x_2) - P_C(x_1), x_2 - x_1 \rangle \geq \|P_C(x_2) - P_C(x_1)\|^2,$
- 2) *monotonicity*: $\forall (x_1, x_2) \in \mathbb{E}^2, \langle P_C(x_2) - P_C(x_1), x_2 - x_1 \rangle \geq 0,$
- 3) *contraction*: $\forall (x_1, x_2) \in \mathbb{E}^2, \|P_C(x_1) - P_C(x_2)\| \leq \|x_1 - x_2\|.$

Remark 1.12 The projector P_C is a nonsmooth operator. It is even not guaranteed to be directionally differentiable when the convex set C is arbitrary (see Kruskal [86] for a counter-example in \mathbb{R}^3 and Shapiro [128] for a counter-example in \mathbb{R}^2). However, whatever the convex set C is, P_C has directional derivatives at a point x belonging to C [137]:

$$\forall x \in C, \forall d \in \mathbb{E}: P'_C(x; d) = P_{T_C(x)} d. \quad (1.30)$$

It is also worth noting that the projector P_C is smoother when the boundary of the convex set C is smooth in a sense that is described in [69]. \square

1.2.8 Asymptotic Cone

Let \mathbb{E} be a vector space of *finite dimension* and C be a nonempty closed convex set of \mathbb{E} . The *asymptotic cone* of C is defined and denoted by

$$C^\infty := \{d \in \mathbb{E} : C + \mathbb{R}_+ d \subseteq C\}.$$

This cone has also the following expressions, whatever $x \in C$ is:

$$\begin{aligned} C^\infty &= \{d \in \mathbb{E} : C + d \subseteq C\} \\ &= \{d \in \mathbb{E} : x + \mathbb{R}_+ d \subseteq C\} = \bigcap_{t>0} \frac{C - x}{t} \\ &= \left\{ d \in \mathbb{E} : \exists \{x_k\} \subseteq C, \exists \{t_k\} \rightarrow \infty \text{ such that } \frac{x_k}{t_k} \rightarrow d \right\}. \end{aligned} \quad (1.31)$$

The formula (1.31) shows that C^∞ is closed.

The asymptotic cone is a nice tool to determine, by calculation, whether a closed convex set is bounded.

Proposition 1.13 (boundedness by calculation) *Let C be a nonempty closed convex set. Then,*

$$C \text{ is bounded} \iff C^\infty = \{0\}.$$

Proposition 1.14 (asymptotic cone calculus)

- 1) If $K \neq \emptyset$, then K is a closed convex cone if and only if $K^\infty = K$.
- 2) For an arbitrary collection $\{C_i\}_{i \in I}$ of closed convex sets C_i with nonempty intersection:

$$(\cap_{i \in I} C_i)^\infty = \cap_{i \in I} C_i^\infty.$$

- 3) Let $A : \mathbb{E} \rightarrow \mathbb{F}$ and $B : \mathbb{E} \rightarrow \mathbb{G}$ be linear maps, $a \in \mathbb{F}$, $b \in \mathbb{G}$, K be a nonempty closed convex cone of \mathbb{G} , and

$$P := \{x \in \mathbb{E} : Ax = a, Bx \in b + K\} \neq \emptyset.$$

Then,

$$P^\infty = \{d \in \mathbb{E} : Ad = 0, Bd \in K\}.$$

Let \mathbb{E} be a Euclidean space. The sets S_1 and $S_2 \subseteq \mathbb{E}$ are said to be *strictly separable* if there exists a vector $\xi \in \mathbb{E}$ (necessarily nonzero) such that

$$\sup_{x_1 \in S_1} \langle \xi, x_1 \rangle < \inf_{x_2 \in S_2} \langle \xi, x_2 \rangle.$$

Proposition 1.15 (strict separation of convex sets) One can strictly separate two disjoint nonempty closed convex sets C_1 and C_2 of a Euclidean space \mathbb{E} in any of the following situations

- 1) $C_1 - C_2$ is closed,
- 2) $C_1^\infty \cap C_2^\infty = \{0\}$,
- 3) C_1 or C_2 is compact,
- 4) C_1 and C_2 are polyhedral.

1.2.9 Convex Function

Let \mathbb{E} be a vector space. The *domain* and the *epigraph* of an arbitrary (not necessarily convex) function $f : \mathbb{E} \rightarrow \overline{\mathbb{R}}$ are the sets defined and denoted by

$$\begin{aligned} \text{dom } f &:= \{x \in \mathbb{E} : f(x) < +\infty\}, \\ \text{epi } f &:= \{(x, \alpha) \in \mathbb{E} \times \mathbb{R} : f(x) \leq \alpha\}. \end{aligned}$$

The *indicator function* of an arbitrary (not necessarily convex) set $S \subseteq \mathbb{E}$ is the function $\mathcal{I}_S : \mathbb{E} \rightarrow \mathbb{R} \cup \{+\infty\}$ defined at $x \in \mathbb{E}$ by

$$\mathcal{I}_S(y) = \begin{cases} 0 & \text{if } y \in S, \\ +\infty & \text{otherwise.} \end{cases}$$

By definition, a function $f : \mathbb{E} \rightarrow \overline{\mathbb{R}}$ is *convex* if its epigraph is convex; it is *closed* if its epigraph is closed in $\mathbb{E} \times \mathbb{R}$; it is *proper* if it does not take the value $-\infty$ and is not identical to $+\infty$. The set of proper convex functions $f : \mathbb{E} \rightarrow \overline{\mathbb{R}}$ is denoted by

$$\text{Conv}(\mathbb{E})$$

and the set of closed proper convex functions $f : \mathbb{E} \rightarrow \overline{\mathbb{R}}$ is denoted by

$$\overline{\text{Conv}}(\mathbb{E}).$$

A function $f : \mathbb{E} \rightarrow \overline{\mathbb{R}}$ is said to be *directionally differentiable* in the direction $d \in \mathbb{E}$ at a point $x \in \mathbb{E}$ at which it is finite if the following limit exists in $\overline{\mathbb{R}}$:

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

The value $f'(x; d) \in \overline{\mathbb{R}}$ is then called the *directional derivative* of f at x in the direction d . A convex function always has directional derivatives, but these can take infinite values.

Proposition 1.16 (directional differentiability) *Let $f : \mathbb{E} \rightarrow \overline{\mathbb{R}}$ be a convex function, $x \in \mathbb{E}$ be a point at which $f(x)$ is finite, and $d \in \mathbb{E}$. Then,*

1) *the function*

$$t \in \mathbb{R}_{++} \mapsto \frac{f(x + td) - f(x)}{t} \in \overline{\mathbb{R}}$$

est nondecreasing,

- 2) *$f'(x; d)$ exists in $\overline{\mathbb{R}}$ (it can take the values $-\infty$ or $+\infty$),*
 3) *$f'(x; d) = +\infty$ if and only if $x + td \notin \text{dom } f$ for all $t > 0$,*
 4) *there holds*

$$f'(x; d) \geq -f'(x; -d), \tag{1.32}$$

in particular, if one of the two directional derivatives $f'(x; d)$ or $f'(x; -d)$ is $-\infty$ the other is $+\infty$.

Remarks 1.17 1) According to point 3, one has $f'(x; d) = +\infty$ if and only if $f(x + td) = +\infty$ for all $t > 0$. but, one can very well have $f'(x; d) = -\infty$, while $f(x + td) \geq -\infty$ for all $t > 0$. This is the case at $x = 0$ for the convex function $f : \mathbb{R} \rightarrow \overline{\mathbb{R}}$ defined at $x \in \mathbb{R}$ by

$$f(x) = \begin{cases} -\sqrt{x} & \text{if } x \geq 0 \\ +\infty & \text{otherwise} \end{cases} \quad \text{and} \quad \begin{array}{c} \begin{array}{c} 0 \\ | \\ \hline \end{array} \\ \begin{array}{c} \text{---} \\ \diagdown \\ \text{---} \end{array} \\ \begin{array}{c} -\sqrt{x} \end{array} \end{array} \tag{1.33}$$

and the direction $d = 1$.

- 2) Point 4 shows that one can compare $f'(x; d)$ and $f'(x; -d)$ when the function f is convex. If f is not differentiable at x , in general $f'(x; d) \neq -f'(x; -d)$. For example, if $f(x) = |x|$, $x \in \mathbb{R}$, one has $f'(0; 1) = f'(0; -1) = 1$.

Function (1.33) allows us to see what formula (1.32) yields with infinite directional derivatives: $f'(0; 1) = -\infty$ implies that $f'(0; -1) = +\infty$ must hold.

1.2.10 Asymptotic Function

The epigraph of a function $f \in \overline{\text{Conv}}(\mathbb{E})$ is a nonempty closed convex set. One can therefore consider its **asymptotic cone** $(\text{epi } f)^\infty$. This one has interesting properties.

Proposition 1.18 (asymptotic function f^∞) *If $f \in \overline{\text{Conv}}(\mathbb{E})$, then*

- 1) $(\text{epi } f)^\infty$ is the epigraph of a function $f^\infty : \mathbb{E} \rightarrow \mathbb{R} \cup \{+\infty\}$,
- 2) for all $x \in \text{dom } f$ and all $d \in \mathbb{E}$

$$f^\infty(d) = \lim_{t \rightarrow \infty} \frac{f(x + td) - f(x)}{t} = \lim_{t \rightarrow \infty} \frac{f(x + td)}{t}, \quad (1.34)$$

- 3) $\text{dom } f^\infty \subseteq (\text{dom } f)^\infty$,
- 4) $f^\infty \in \overline{\text{Conv}}(\mathbb{E})$.

Proof. 0) Let us start by giving a characterization of the membership to $(\text{epi } f)^\infty$. For this, take an arbitrary point $(x, f(x))$ in $\text{epi } f$, which is nonempty when $f \in \overline{\text{Conv}}(\mathbb{E})$. Then,

$$\begin{aligned} (d, \delta) \in (\text{epi } f)^\infty &\iff (x, f(x)) + t(d, \delta) \in \text{epi } f, \quad \forall t > 0 \\ &\iff f(x + td) \leq f(x) + t\delta, \quad \forall t > 0. \end{aligned} \quad (1.35)$$

1) To be an epigraph, $(\text{epi } f)^\infty$ must have two properties.

- The first property is that, when $(d, \delta) \in (\text{epi } f)^\infty$ and $\delta' \geq \delta$, it must follow that $(d, \delta') \in (\text{epi } f)^\infty$. This is actually clear by (1.35).
- The second property is that for any $d \in \mathbb{E}$ such that $\{d\} \times \mathbb{R}$ intersects $(\text{epi } f)^\infty$, one must have $(d, \delta_0) \in (\text{epi } f)^\infty$ for

$$\delta_0 := \inf\{\delta : (d, \delta) \in (\text{epi } f)^\infty\}. \quad (1.36a)$$

From (1.35), we get

$$\delta_0 = \sup_{t > 0} \frac{f(x + td) - f(x)}{t}. \quad (1.36b)$$

By definition of the supremum, one has then $f(x + td) \leq f(x) + t\delta_0$, for all $t > 0$. This and (1.35) now yield that $(d, \delta_0) \in (\text{epi } f)^\infty$.

Denote by f^∞ the function whose epigraph is $(\text{epi } f)^\infty$.

2) Let $x \in \text{dom } f$ and $d \in \mathbb{E}$. It is clear that $(x, f(x)) \in \text{epi } f$.

Let us show that $f^\infty(d)$ is the value δ_0 given by (1.36a), hence by (1.36a), which will prove the first equality in (1.34). By the reasoning of the previous point, $f^\infty(d)$ is clearly δ_0 when $\{d\} \times \mathbb{R}$ intersects $(\text{epi } f)^\infty$. When $\{d\} \times \mathbb{R}$ does not intersect $(\text{epi } f)^\infty = \text{epi } f^\infty$, both $f^\infty(d)$ and δ_0 take the value $= +\infty$ (the first one by definition of the epigraph of f^∞ , the second one by definition of the infimum in (1.36a) and the fact that there is no $(d, \delta) \in (\text{epi } f)^\infty$) hence $f^\infty(d) = \delta_0$ in that case also.

For the second equality in (1.34), just observe that $f(x) \in \mathbb{R}$ because $x \in \text{dom } f$ and $f \in \overline{\text{Conv}}(\mathbb{E})$.

3) If $d \in \text{dom } f^\infty$, then $(d, f^\infty(d)) \in (\text{epi } f)^\infty$ and

$$f(x + td) \leq f(x) + tf^\infty(d), \quad \forall t > 0,$$

by (1.35). Since the right-hand side of this inequality is finite for all positive t , it results that $d \in (\text{dom } f)^\infty$.

4) First, f^∞ is a closed convex function, since its epigraph is the closed convex set $(\text{epi } f)^\infty$. Secondly, $f^\infty \not\equiv +\infty$, since its epigraph $(\text{epi } f)^\infty$ is nonempty when $f \in \overline{\text{Conv}}(\mathbb{E})$. Thirdly, $f > -\infty$ since, by (1.34), for $d \in \text{dom } f^\infty$, $f(d)$ is obtained as a limit of an increasing sequence made of finite values. \square

Remarks 1.19 1) The *differential quotient* $[f(x + td) - f(x)]/t$ is nondecreasing with t . It converges to $f'(x; d)$ when $t \downarrow 0$ and, according to (1.34), it converges to $f^\infty(d)$ when $t \rightarrow \infty$.

- 2) the last identity in (1.34) often provides the easiest way of computing $f^\infty(d)$. Note however that, unlike the differential quotient, $f(x + td)/t$ is not monotonic with t .
- 3) The inclusion in point 3 may be strict. For the exponential function $\exp \in \overline{\text{Conv}}(\mathbb{R}) : x \mapsto \exp(x) = e^x$, there holds $\exp^\infty(1) = +\infty$, so that $1 \notin \text{dom}(\exp^\infty)$. But $1 \in (\text{dom } \exp)^\infty$, since $\text{dom } \exp = \mathbb{R}$.
- 4) The inclusion in point 3 expresses in a compact manner the fact that if $f(x + td) = +\infty$ for some $t > 0$, then $f^\infty(d) = +\infty$. \square

The *sublevel set* of an arbitrary function $f : \mathbb{E} \rightarrow \mathbb{R} \cup \{+\infty\}$ of level $\alpha \in \mathbb{R}$ is the set

$$L_\alpha(f) := \{x \in \mathbb{E} : f(x) \leq \alpha\}.$$

Whilst the asymptotic cone is a useful concept to verify the boundedness of a nonempty closed convex set, the asymptotic function is useful to verify the boundedness of the sublevel sets of the corresponding function, in particular the nonemptiness and boundedness of the set of its minimizers.

Proposition 1.20 (existence of a bounded set of minimizers) *If $f \in \overline{\text{Conv}}(\mathbb{E})$, then*

- 1) $\forall \alpha \in \mathbb{R}$ such that $L_\alpha(f) \neq \emptyset$, the following holds

$$[L_\alpha(f)]^\infty = \{d \in \mathbb{E} : f^\infty(d) \leq 0\}, \quad (1.37)$$

- 2) the following properties are equivalent:

- (i) $\exists \alpha \in \mathbb{R} : L_\alpha(f)$ is nonempty and bounded,
- (ii) $\forall \alpha \in \mathbb{R} : L_\alpha(f)$ is bounded,
- (iii) $\text{Arg min } f$ is nonempty and bounded,
- (iv) $\forall d \in \mathbb{E} \setminus \{0\} : f^\infty(d) > 0$.

Proof. 1) Let $\alpha \in \mathbb{R}$ and $x \in L_\alpha(f)$, which is assumed to be nonempty. Then,

$$d \in (L_\alpha(f))^\infty \iff f(x + td) \leq \alpha, \quad \forall t \geq 0.$$

Hence, if $d \in (L_\alpha(f))^\infty$, $f^\infty(d) \leq 0$ by using the last identity in (1.34). Conversely, if $f^\infty(d) \leq 0$, then $f(x + td) \leq f(x)$, for all $t > 0$, by the monotonicity of the differential quotient and (1.34). Since $f(x) \leq \alpha$, $x \in L_\alpha(f)$, this last inequality yields $f(x + td) \leq \alpha$, for all $t > 0$, which shows that $x + td \in L_\alpha(f)$, for all $t > 0$; hence, $d \in (L_\alpha(f))^\infty$.

2) Let us now consider the equivalences.

[(i) \Rightarrow (ii)] By (i) and proposition 1.13, it results that $[L_\alpha(f)]^\infty = \{0\}$ and therefore $\{d \in \mathbb{E} : f^\infty(d) \leq 0\} = \{0\}$ by (1.37). Since $\{d \in \mathbb{E} : f^\infty(d) \leq 0\}$ is independent of α , it results that $(L_{\alpha'}(f))^\infty = \{0\}$ for all α' such that $L_{\alpha'}(f) \neq \emptyset$. Hence, (ii) holds.

[(ii) \Rightarrow (iii)] Since f is proper, one can find an α such that $L_\alpha(f) \neq \emptyset$. Then $L_\alpha(f)$ is nonempty and closed (since $f \in \text{Conv}(\mathbb{E})$); it is also compact by (ii). As a result, the function f reaches its minimum at some point $\bar{x} \in L_\alpha(f)$; this point \bar{x} is also a minimum of f on \mathbb{E} . Since $\text{Arg min } f = L_\alpha(f)$ with $\alpha = \min f$, $\text{Arg min } f$ is compact by (ii).

[(iii) \Rightarrow (i)] Clear with $\alpha = \min f$.

[(i) \Rightarrow (iv)] By (i), it results that $(L_\alpha(f))^\infty = \{0\}$ and therefore that $\{d \in \mathbb{E} : f^\infty(d) \leq 0\} = \{0\}$ by (1.37). This implies (iv).

[(iv) \Rightarrow (ii)] By (iv), $\{d \in \mathbb{E} : f^\infty(d) \leq 0\} = \{0\}$ and therefore $(L_\alpha(f))^\infty = \{0\}$ each time $L_\alpha(f) \neq \emptyset$ (by using (1.37)). One deduces (ii). \square

The implication (iv) \Rightarrow (iii) of proposition 1.20 is often a convenient approach to show that a function $f \in \text{Conv}(\mathbb{E})$ has a nonempty and bounded set of minimizers (it is ineffective if the set of minimizers is unbounded). The technique consists, therefore, in calculating the asymptotic function f^∞ and in highlighting its property (iv). Hence, showing the existence a nonempty compact set of minimizers is a task that can be realized by calculation.

1.2.11 Subdifferentiability

The notion of subdifferentiability of a function $f \in \text{Conv}(\mathbb{E})$ is based on the following proposition.

Proposition 1.21 (subdifferentiability) *Suppose that $f \in \text{Conv}(\mathbb{E})$, $x \in \text{dom } f$ and $x^* \in \mathbb{E}$. Then, the following properties are equivalent:*

- (i) $\forall d \in \mathbb{E} : f'(x; d) \geq \langle x^*, d \rangle$,
- (ii) $\forall y \in \mathbb{E} : f(y) \geq f(x) + \langle x^*, y - x \rangle$,
- (iii) $x \in \text{Arg min}_{y \in \mathbb{E}} (f(y) - \langle x^*, y \rangle) = \text{Arg max}_{y \in \mathbb{E}} (\langle x^*, y \rangle - f(y))$.

Definitions 1.22 A function $f \in \text{Conv}(\mathbb{E})$ is said to be *subdifferentiable* at a point $x \in \text{dom } f$ if there exists $x^* \in \mathbb{E}$ verifying the equivalent properties (i)-(iii) of proposition 1.21. Such an element x^* is called a *subgradient* of f at x . The set of subgradients of f at x is called the *subdifferential* of f at x and is denoted by $\partial f(x)$. By convention, $\partial f(x) = \emptyset$ if $x \notin \text{dom } f$. \square

A subgradient is actually an element of the dual of \mathbb{E} , which is identified to \mathbb{E} itself in finite dimension (the distinction between \mathbb{E} and its dual is important in infinite dimension but not in our finite dimension framework). Hence, it is better to see a subgradient as a *slope* of the function, like for a gradient of a smooth function.

Each of the three equivalent conditions (i)-(iii) of proposition 1.21 reflects a particular aspect of the subgradients, offering also a means to calculate them.

- 1) Condition (i) tells us that a subgradient is the slope of *linear* function minorizing $f'(x; \cdot)$. This point of view leads to the following way of computing $\partial f(x)$. First, one computes the directional derivatives of f at x and one determines all the linear function minorizing f ; their slopes are the elements of $\partial f(x)$.
- 2) According to condition (ii), a subgradient is the slope of an *affine* function minorizing f , which has the same value as f at x . This definition is often used to verify that a particular $x^* \in \mathbb{E}$ is a subgradient.
- 3) Condition (iii) tells us that x^* is a subgradient of f at x if $f(\cdot) - \langle x^*, \cdot \rangle$ reaches its minimum at x . This point of view leads to the following method to compute $\partial f(x)$. One starts with a slope $x^* \in \mathbb{E}$; then, one computes the minimizers of $x \mapsto f(x) - \langle x^*, x \rangle$; such a minimizer x is such that $x^* \in \partial f(x)$.

A function $f \in \text{Conv}(\mathbb{E})$ is not necessarily subdifferentiable at all the points in $\text{dom } f$. For example, the function (1.33) is not subdifferentiable at 0.

Notes

Most results given in this section can be found with their proofs in [125, 66, 22, 67]. The notions of **asymptotic cone** and function have been extended to nonconvex sets and functions, in particular with the goal of providing existence theorems for nonconvex optimization problems and variational inequalities; see the monograph [7].

Exercises

- 1.2.1.** *Constant subsequence of a sequence of subsets of $[1:n]$.* Let $\{I_k\}_{k \in \mathbb{N}}$ be a sequence of subsets I_k of $[1:n]$. Show that this sequence contains a subsequence $\{I_k\}_{k \in \mathcal{K}}$ (hence \mathcal{K} is an infinite part of \mathbb{N}) such that I_k is independent of $k \in \mathcal{K}$.
- 1.2.2.** *Affine hull.* Let A be an affine subspace of a vector space \mathbb{E} and $O \subseteq A$ be relatively open in A (i.e., open for the relative topology of A). Then, $\text{aff } O = A$.
- 1.2.3.** *Scaled sum of convex sets.* If C_1 and C_2 are convex sets of a vector space \mathbb{E} , such that $0 \in C_1 \cap C_2$, then $\mathbb{R}_+(C_1 + C_2) = \mathbb{R}_+C_1 + \mathbb{R}_+C_2$. Give an example, in which the identity does not hold when $0 \notin C_1 \cap C_2$.
- 1.2.4.** *Dense convex set.* Let C be a convex set in a finite dimensional vector space \mathbb{E} . Show that if the closure of C is \mathbb{E} , then $C = \mathbb{E}$.
- 1.2.5.** *Relative interior.* Let C be a nonempty convex set of a finite dimensional vector space. Then,

$$2(\text{ri } C) = C + \text{ri } C = \overline{C} + \text{ri } C.$$

- 1.2.6.** *Affine hull and relative interior of a convex polyhedron.* Let \mathbb{E} be a vector space. Consider the convex polyhedron of \mathbb{E} defined by

$$P := \{x \in \mathbb{E} : Ax = a, Bx \leq b\},$$

where $A : \mathbb{E} \rightarrow \mathbb{R}^m$ and $B : \mathbb{E} \rightarrow \mathbb{R}^p$ are linear maps, $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^p$. Define the index set $I \subseteq [1 : p]$ by

$$i \in I \iff \{x \in P : B_i x < b_i\} \neq \emptyset.$$

and denote its complementary set by $I^c := [1 : p] \setminus I$. Show that

$$\text{aff } P = \{x \in \mathbb{E} : Ax = a, B_{I^c} x = b_{I^c}\}, \quad (1.38)$$

$$\text{ri } P = \{x \in P : B_I x < b_I\}. \quad (1.39)$$

- 1.2.7.** *Decomposition of a vector space in the sum of a subspace and a convex cone* [129]. Let \mathbb{E}_0 be a subspace of a vector space \mathbb{E} and K be a convex cone of \mathbb{E} . Then,

$$\mathbb{E}_0 + K = \mathbb{E} \iff \begin{cases} \mathbb{E}_0 + \text{vect } K = \mathbb{E} \\ \mathbb{E}_0 \cap (\text{ri } K) \neq \emptyset. \end{cases}$$

- 1.2.8.** *Farkas lemma for a polyhedral cone.* Let \mathbb{E} be a Euclidean space, and $A_E : \mathbb{E} \rightarrow \mathbb{R}^{m_E}$ and $A_J : \mathbb{E} \rightarrow \mathbb{R}^{m_J}$ be two linear maps. Then,

$$\{d : A_E d = 0, A_J d \leq 0\}^+ = -\{A_E^* \lambda_E + A_J^* \lambda_J : \lambda_E \in \mathbb{R}^{m_E}, \lambda_J \in \mathbb{R}_+^{m_J}\}. \quad (1.40)$$

- 1.2.9.** *Bidual of a convex cone.* Using Farkas lemma, show that if K is a nonempty convex cone, then $K^{++} = \overline{K}$.
- 1.2.10.** *Interior and relative interior of a dual cone.* Let \mathbb{E} be a Euclidean (scalar product and associated norm respectively denoted by $\langle \cdot, \cdot \rangle$ and $\|\cdot\|$), $P \subseteq \mathbb{E}$, P^+ be the dual cone of P , and $P_{\text{aff}(P^+)}$ be the orthogonal projector on $\text{aff}(P^+)$. Show that

$$d \in \text{int}(P^+) \iff \exists \varepsilon > 0, \forall x \in P: \langle d, x \rangle \geq \varepsilon \|x\|, \quad (1.41a)$$

$$d \in \text{ri}(P^+) \iff \exists \varepsilon > 0, \forall x \in P: \langle d, x \rangle \geq \varepsilon \|P_{\text{aff}(P^+)} x\|. \quad (1.41b)$$

- 1.2.11.** *Cone of feasible directions to a convex cone* [129]. Show that if K is a convex cone and $x \in K$, then $T_x^f K = K + \mathbb{R}\{x\}$.
- 1.2.12.** *Tangent cones to \mathcal{S}_+^n .* Show that the tangent cone to \mathcal{S}_+^n at $X \in \mathcal{S}_+^n$ can be written

$$T_X \mathcal{S}_+^n = \{D \in \mathcal{S}^n : v^T D v \geq 0, \forall v \in \mathcal{N}(X)\}.$$

- 1.2.13.** *Asymptotic cone.* Let \mathbb{E}, \mathbb{F} , and \mathbb{G} be finite dimension vector spaces. Let $A : \mathbb{E} \rightarrow \mathbb{F}$ and $B : \mathbb{E} \rightarrow \mathbb{G}$ be linear maps, $a \in \mathbb{F}$, $b \in \mathbb{G}$, $K \subseteq \mathbb{G}$ be a closed convex cone, and

$$P = \{x \in \mathbb{E} : Ax = a, Bx \in b + K\}.$$

Then, the asymptotic cone of P is given by

$$P^\infty = \{d \in \mathbb{E} : Ad = 0, Bd \in K\}.$$

1.3 Nonsmooth Analysis

1.3.1 Lower semi-continuity

Let \mathbb{E} be a finite dimensional vector space and $f : \mathbb{E} \rightarrow \overline{\mathbb{R}}$ be a function. The function f is said to be *lower semi-continuous* (*l.s.c.* for short) on \mathbb{E} if for all $x \in \mathbb{E}$ there holds

$$f(x) \leq \liminf_{x' \rightarrow x} f(x').$$

The function f is said to be *closed* if its epigraph is closed. These two notions are actually equivalent.

Proposition 1.23 (l.s.c. function) *Let \mathbb{E} be a finite dimensional vector space and $f : \mathbb{E} \rightarrow \overline{\mathbb{R}}$ be a function. Then, the following properties are equivalent.*

- (i) f is l.s.c. on \mathbb{E} ,
- (ii) f is closed,
- (iii) for all $\alpha \in \mathbb{R}$, the sublevel set $L_\alpha(f)$ is closed.

1.3.2 Lipschitz Continuity

Let \mathbb{E} and \mathbb{F} be two finite dimensional normed spaces and $F : \mathbb{E} \rightarrow \mathbb{F}$ be a function. We adopt the following definitions.

- F is *Lipschitz (continuous)* on a set $V \subseteq \mathbb{E}$ if

$$\exists L \geq 0, \quad \forall (x, x') \in V^2 : \quad \|F(x) - F(x')\| \leq L\|x - x'\|.$$

In this case, we also say that F is L -Lipschitz on V .

- F is *radially Lipschitz (continuous)* at $x \in \mathbb{E}$ if

$$\exists V \in \mathcal{N}(x), \quad \exists L \geq 0, \quad \forall x' \in V : \quad \|F(x) - F(x')\| \leq L\|x - x'\|,$$

where $\mathcal{N}(x)$ denotes the family of neighborhoods of x .

- F is *Lipschitz (continuous) near $x \in \mathbb{E}$* if F is Lipschitz on some neighborhood of x .
- F is *locally Lipschitz (continuous) on an open set $\Omega \subseteq \mathbb{E}$* if F is Lipschitz near any point of Ω .

1.3.3 Differentiability

Let \mathbb{E} and \mathbb{F} be two normed spaces and Ω be an open set of \mathbb{E} . A function $F : \Omega \rightarrow \mathbb{F}$ is said to be *Fréchet-differentiable* [57; 1911] at $x \in \Omega$ if there exists a linear continuous operator $L : \mathbb{E} \rightarrow \mathbb{F}$ such that

$$\lim_{\|h\| \downarrow 0} \frac{1}{\|h\|} (F(x+h) - F(x) - Lh) = 0. \quad (1.42)$$

Instead of saying Fréchet-differentiable, one also say *F-differentiable* or simply *differentiable*. The limit in (1.42) is in \mathbb{F} . We have taken care of having the limit for $h \neq 0$ to give a sense to the quotient in (1.42). The operator L is then uniquely determined, is denoted $F'(x)$, and is called the *derivative* of F at x . Condition (1.42) can then also be written

$$F(x+h) = F(x) + F'(x)h + o(h), \quad (1.43)$$

where the *small o* of h , $o(h)$, denotes a function of h vanishing at zero and verifying the property

$$\lim_{\|h\|_{\mathbb{E}} \downarrow 0} \frac{o(h)}{\|h\|} = 0,$$

where the limit is taken in a normed space depending on the context (here \mathbb{F}).

The function $F : \Omega \rightarrow \mathbb{F}$ is said to be *F-differentiable on Ω* if F is F-differentiable at each point of Ω .

Theorem 1.24 (mean value theorem) *Let Ω be an open set of \mathbb{E} , $x \in \Omega$, and $h \in \mathbb{E} \setminus \{0\}$ be such that the closed segment $[x, x+h] \subseteq \Omega$. Suppose that $F : \Omega \rightarrow \mathbb{F}$ is continuous on Ω and differentiable on the open segment $(x, x+h)$. Then*

$$\|F(x+h) - F(x)\| \leq \left(\sup_{z \in (x, x+h)} \|F'(z)\| \right) \|h\|.$$

Corollary 1.25 *Under the assumptions of theorem 1.24, if $L : \mathbb{E} \rightarrow \mathbb{F}$ is linear continuous, it follows that*

$$\|F(x+h) - F(x) - Lh\| \leq \left(\sup_{z \in (x, x+h)} \|F'(z) - L\| \right) \|h\|.$$

Suppose that $F : \Omega \rightarrow \mathbb{F}$ is defined on the part $\Omega \subseteq \mathbb{E}$. We denote by \mathcal{D}_F the set of points of Ω at which F is Fréchet-differentiable:

$$\mathcal{D}_F := \{x \in \Omega : F \text{ is Fréchet-differentiable at } x\}.$$

Theorem 1.26 (Rademacher, 1919) *If Ω is an open set and $F : \Omega \rightarrow \mathbb{F}$ is locally Lipschitz on Ω , then, F is Fréchet-differentiable almost everywhere on Ω , in the sense of Lebesgue. In other words, the Lebesgue measure of $\Omega \setminus \mathcal{D}_F$ vanishes.*

Proof. See [49; 2015, chap. 3] for instance. □

1.3.4 Multifunction

A *multifunction*³ T between two sets \mathbb{E} and \mathbb{F} is a usual function from \mathbb{E} to $\mathcal{P}(\mathbb{F})$, where $\mathcal{P}(\mathbb{F})$ is the *power set* of \mathbb{F} , that is the set of all subsets of \mathbb{F} . The adopted notation for a multifunction is

³ A multifunction may have many other names, like *set-valued mapping* or *multi-valued function*.

$$T : \mathbb{E} \multimap \mathbb{F} : x \in \mathbb{E} \mapsto T(x) \subseteq \mathbb{F}.$$

Several concepts are associated with a multifunction $T : \mathbb{E} \multimap \mathbb{F}$.

- The *graph* $\mathcal{G}(T)$, the *domain* $\mathcal{D}(T)$, and the *range* $\mathcal{R}(T)$ of T are defined by

$$\begin{aligned} \mathcal{G}(T) &:= \{(x, y) \in \mathbb{E} \times \mathbb{F} : y \in T(x)\}, \\ \mathcal{D}(T) &:= \{x \in \mathbb{E} : (x, y) \in \mathcal{G}(T) \text{ for some } y \in \mathbb{F}\} = \pi_{\mathbb{E}}\mathcal{G}(T), \\ \mathcal{R}(T) &:= \{y \in \mathbb{F} : (x, y) \in \mathcal{G}(T) \text{ for some } x \in \mathbb{E}\} = \pi_{\mathbb{F}}\mathcal{G}(T), \end{aligned}$$

where we have denoted by $\pi_{\mathbb{E}} : (x, y) \in \mathbb{E} \times \mathbb{F} \mapsto x \in \mathbb{E}$ and $\pi_{\mathbb{F}} : (x, y) \in \mathbb{E} \times \mathbb{F} \mapsto y \in \mathbb{F}$, the *Cartesian projectors* on \mathbb{E} and \mathbb{F} respectively (these are linear maps). Notice that $\mathcal{G}(T)$ is a part of $\mathbb{E} \times \mathbb{F}$, not of $\mathbb{E} \times \mathcal{P}(\mathbb{F})$.

- The concept of multifunction is the same as the one of *binary relation*, which is defined by the specification of a part G of $\mathbb{E} \times \mathbb{F}$ (the relation, sometimes denoted $x \mathcal{R} y$, is said to be “true” if $(x, y) \in G$ and “false” otherwise). Indeed, one can associate with a part G of $\mathbb{E} \times \mathbb{F}$ the multifunction $T_G : \mathbb{E} \multimap \mathbb{F}$, defined at $x \in \mathbb{E}$ by

$$T_G(x) = \{y \in \mathbb{F} : (x, y) \in G\}.$$

Then, we have the property that $\mathcal{G}(T_G) = G$.

- The *image* of a part $P \subseteq \mathbb{E}$ by T is

$$T(P) := \bigcup_{x \in P} T(x) = \pi_{\mathbb{F}}[\mathcal{G}(T) \cap (P \times \mathbb{F})]. \quad (1.44)$$

- The *inverse* of T is the multifunction $T^{-1} : \mathbb{F} \multimap \mathbb{E}$ defined at $y \in \mathbb{F}$ by

$$T^{-1}(y) := \{x \in \mathbb{E} : y \in T(x)\}.$$

The inverse always exists, but it is not true that $T^{-1} \circ T$ is the identity on \mathbb{E} or that $T \circ T^{-1}$ is the identity on \mathbb{F} ! Note that

$$y \in T(x) \iff x \in T^{-1}(y) \iff (x, y) \in \mathcal{G}(T) \iff (y, x) \in \mathcal{G}(T^{-1}).$$

Here are some commonly encountered properties that a multifunction $T : \mathbb{E} \multimap \mathbb{F}$ may have.

- When \mathbb{E}, \mathbb{F} are vector spaces, T is said to be *convex* if $\mathcal{G}(T)$ is convex in $\mathbb{E} \times \mathbb{F}$. This is equivalent to saying that $\forall (x_0, x_1) \in \mathbb{E}^2$ and $\forall t \in [0, 1]$:

$$T((1-t)x_0 + tx_1) \supseteq (1-t)T(x_0) + tT(x_1).$$

Note that

$$\left. \begin{array}{l} T : \mathbb{E} \multimap \mathbb{F} \text{ convex} \\ C \text{ convex in } \mathbb{E} \end{array} \right\} \implies T(C) \text{ convex in } \mathbb{F}. \quad (1.45)$$

This is because by (1.44) there holds $T(C) = \pi_{\mathbb{F}}[\mathcal{G}(T) \cap (C \times \mathbb{F})]$ and $\mathcal{G}(T)$ is convex when T is convex.

- When \mathbb{E}, \mathbb{F} are topological spaces, T is said to be
 - *closed at $x \in \mathbb{E}$* if $y \in T(x)$ when $(x_k, y_k) \in \mathcal{G}(T)$ converges to (x, y) ;
 - *closed* if $\mathcal{G}(T)$ is closed in $\mathbb{E} \times \mathbb{F}$, which amounts to saying that T is closed at any $x \in \mathbb{E}$.
- When \mathbb{E}, \mathbb{F} are metric spaces, T is said to be *upper semi-continuous at $x \in \mathbb{E}$* if

$$\forall \varepsilon > 0, \exists \delta > 0, \forall x' \in x + \delta B : T(x') \subseteq T(x) + \varepsilon B.$$

In this definition, B may be the open or closed unit ball at any place.

Notes

A study of Lipschitz continuous functions can be found in [49; 2015, chap. 3], for functions between finite dimensional vector spaces, including a proof of the Rademacher theorem [115; 1919], and in [4; 2004, chap. 3], between metric spaces. The results on the generalized differentiability are taken from [32, 33].

Exercises

- 1.3.1.** *Upper semi-continuity of some multifunctions.* Let \mathbb{E} be a Euclidean vector space. Show the upper semi-continuity of the following multifunctions.
- 1)
 - 2) The subdifferential of a convex function $f : \mathbb{E} \rightarrow \mathbb{R}$ at $x \in \mathbb{E}$ is the set $\partial f(x) := \{s \in \mathbb{E} : f(y) \geq f(x) + \langle s, y - x \rangle, \forall y \in \mathbb{E}\}$. Show that the multifunction $\partial f : \mathbb{E} \rightarrow \mathbb{E} : x \mapsto \partial f(x)$ is upper semi-continuous at any $x \in \mathbb{E}$.

1.4 Optimization

1.4.1 Generic Problem

In a rather general setting, an optimization problem consists in minimizing a function $f : \mathbb{E} \rightarrow \mathbb{R}$ (\mathbb{E} is a Euclidean vector space, which is the only restriction) on a (possibly nonconvex) subset $X \subseteq \mathbb{E}$. In other words, one looks for a point $x_* \in \mathbb{E}$ such that

$$\begin{cases} x_* \in X, \\ f(x_*) \leq f(x), \quad \forall x \in X. \end{cases} \tag{1.46}$$

Such a point x_* is called a *solution* to the optimization problem. This problem, denoted (P_X) is written in one of the following manners

$$\left\{ \begin{array}{l} \min_{x \in X} f(x) \end{array} \right. \quad \text{or} \quad \inf_{x \in X} f(x) \quad \text{or} \quad \inf \{f(x) : x \in X\}. \tag{1.47}$$

The function f is often called the *objective* of the optimization problem, while the set X is called its *feasible set*. A point belonging to X is said to be *feasible*. The “smallest value” of f on X , more precisely

$$\text{val}(P_X) := \inf_{x \in X} f(x)$$

is called the *optimal value* of the optimization problem, while the set of solutions is denoted by

$$\text{Sol}(P_X) \quad \text{or} \quad \underset{x \in X}{\text{Arg min}} f(x).$$

One says that the problem (P_X) is *bounded* if $\text{val}(P_X) > -\infty$; otherwise, the problem is said to be *unbounded*, which occurs when there is a sequence $\{x_k\} \subseteq X$ such that $f(x_k) \rightarrow -\infty$. We adopt the following convention

$$\inf_{x \in \emptyset} f(x) = +\infty. \quad (1.48)$$

A *minimizing sequence* for problem (P_X) is a sequence $\{x_k\}_{k \geq 1} \subseteq X$ such that $f(x_k) \rightarrow \text{val}(P_X)$ when $k \rightarrow \infty$. By definition of the infimum, such a sequence always exists when X is nonempty.

Maximizing f is “identical” to minimizing $-f$ (i.e., same solutions, opposite optimal value), so that only the minimization problem will be considered. Furthermore, one prefers minimization to maximization, because the notion of “convex set” is meaningful (unlike the one of “concave set”), hence the notion of “convex function” (it is a function whose epigraph is convex), and finally it is more natural to minimize a convex function than to maximize it. From (1.48), one gets

$$\sup_{x \in \emptyset} f(x) = -\infty.$$

A point x_* satisfying (1.46) is sometimes called a *global minimum*, to stress the distinction with a *local minimum* of f on X , which is a point x_* for which there is a neighborhood V of x_* , such that

$$\begin{cases} x_* \in X, \\ f(x_*) \leq f(x), \quad \forall x \in X \cap V. \end{cases} \quad (1.49)$$

One also uses the notion of *strict local/global minimum*, which is a point x_* verifying (1.46)/(1.49) with a strict inequality $f(x_*) < f(x)$ when $x \neq x_*$.

1.4.2 Peano-Kantorovich Optimality Condition

Let \mathbb{E} be a Euclidean space. When x_* minimizes a function f on $X = \mathbb{E}$ and when f is differentiable at x_* , it is known that

$$f'(x_*) = 0.$$

Such a property of the minimizer x_* is called a necessary condition of optimality (“necessary” since it is implied by the optimality of x_*) of the first order (since it only involves the first derivative of f); this property is abbreviated in NC1. Use the gradient of f , this can also be written

$$\nabla f(x_*) = 0,$$

which is sometimes called the *Fermat optimality condition*.

When $X \neq \mathbb{E}$, some kind of first order approximation of X near x_* is also necessary to get a necessary condition of optimality of the first order. The linearization of X at x_* yields a cone that is called the tangent cone. Note that here X is not necessarily convex like in section 1.2.6.

Tangent Cone to a Nonconvex Set

Let X be a *closed* set of \mathbb{E} . A direction $d \in \mathbb{E}$ is said to be *tangent to X at $x \in X$* (in the sense of Bouligand) if

$$\exists \{x_k\} \subseteq X, \quad \exists \{t_k\} \downarrow 0 : \quad \frac{x_k - x}{t_k} \rightarrow d.$$

The *tangent cone* to X at x (in the sense of Bouligand) is the set of tangent directions. It is denoted by

$$T_x X \quad \text{or} \quad T_X(x),$$

with the convention that $T_x X = \emptyset$ when $x \notin X$.

For $x \in X$, one has

$$T_x X \text{ is a closed cone,}$$

$$X \text{ is convex} \implies T_x X \text{ is convex and } T_x X = \overline{\mathbb{R}_+(X - x)}.$$

Therefore, when X is convex, the notions of tangent cones introduced in section 1.2.6 and here coincide, which justifies the similar notation. When X is not convex, one still have $T_x X \subseteq \overline{\mathbb{R}_+(X - x)}$.

The *normal cone* to X at x is then defined by

$$N_x X := (T_x X)^-, \tag{1.50}$$

with the convention that $N_x X = \emptyset$ when $x \notin X$.

Peano-Kantorovich NC1

The following necessary condition of optimality of the first order (NC1) for the generic problem (P_X) is so important for chapter 2 that we present its simple proof. The result just expresses compactly and at the first order (i.e., using the first derivative) the fact that f is not decreasing along the tangent directions to X at x_* when it is locally minimized on X at $x_* \in X$.

Theorem 1.27 (Peano-Kantorovich NC1) *If x_* is a local minimizer of (P_X) and f is differentiable at x_* , then*

$$\nabla f(x_*) \in (\mathbb{T}_{x_*} X)^+. \quad (1.51)$$

Proof. We have to show that $\langle \nabla f(x_*), d \rangle \geq 0$, for all $d \in \mathbb{T}_{x_*} X$. Let $d \in \mathbb{T}_{x_*} X \setminus \{0\}$ (the previous inequality is trivially satisfied for $d = 0$). Then, there exist sequences $\{x_k\} \subseteq X$ and $\{t_k\} \downarrow 0$ such that $d_k := (x_k - x_*)/t_k \rightarrow d$. For large k , $x_k = x_* + t_k d_k$ is close to x_* , so that, by the local optimality of x_* :

$$f(x_* + t_k d_k) \geq f(x_*), \quad \text{for large } k.$$

Since f is differentiable at x_* , $f(x_* + t_k d_k) = f(x_*) + f'(x_*) \cdot (t_k d_k) + o(\|t_k d_k\|)$. Using the previous inequalities yields

$$0 \leq f'(x_*) \cdot d_k + \frac{o(\|t_k d_k\|)}{t_k} = f'(x_*) \cdot d_k + \frac{o(\|t_k d_k\|)}{\|t_k d_k\|} \|d_k\|, \quad \text{for large } k.$$

Taking the limit when $k \rightarrow \infty$ now provides $f'(x_*) \cdot d_k = \langle \nabla f(x_*), d \rangle \geq 0$. \square

NSC1 for Convex Problems

For a convex problem (P_X) , one has a necessary and sufficient condition of *global* optimality of the first order (NSC1). Recall from proposition 1.16 that a convex function $f : \mathbb{E} \rightarrow \mathbb{R} \cup \{+\infty\}$ has directional derivatives $f'(x; d) := \lim_{t \downarrow 0} [f(x + td) - f(x)]/t \in \overline{\mathbb{R}}$ for all $x \in \text{dom } f$ and all $d \in \mathbb{E}$.

Proposition 1.28 (NSC1 for a convex problem) *Suppose that X is convex, that f is convex on X , and that $x_* \in X$. Then, x_* is a global solution to (P_X) if and only if*

$$f'(x_*; x - x_*) \geq 0, \quad \forall x \in X.$$

The proof is straightforward. It uses the convexity inequality

$$f(x) \geq f(x_*) + f'(x_*; x - x_*), \quad \forall x \in X.$$

1.4.3 Equality Constrained Problem (P_E)

Let \mathbb{E} and \mathbb{F} be Euclidean vector spaces. The *equality constrained problem* consists in minimizing a function $f : \mathbb{E} \rightarrow \mathbb{R}$ on the set

$$X_E := \{x \in \mathbb{E} : c(x) = 0\}$$

defined by equality constraints thanks to a function $c : \mathbb{E} \rightarrow \mathbb{F}$. The set X_E is said to be the *feasible set* of (P_E) . The problem is written

$$(P_E) \quad \begin{cases} \inf_x f(x) \\ c(x) = 0. \end{cases}$$

Usually the functions f and c are smooth, possibly nonconvex.

Definition 1.29 The problem (P_E) is said to be *convex* if f is convex and X_E is a convex set.

Lagrange Optimality Conditions

The following NC1 is often attributed to Lagrange (XVIIIth century). It uses the *Lagrangian* of (P_E) , which is the function

$$\ell : (x, \lambda) \in \mathbb{E} \times \mathbb{F} \mapsto \ell(x, \lambda) = f(x) + \langle \lambda, c(x) \rangle \in \mathbb{R}.$$

We denote by $\nabla_x \ell(x, \lambda)$ the gradient of $\nabla_x \ell(\cdot, \lambda)$ at x . It is a particular case of theorem 1.40, whose proof is given explicitly.

Theorem 1.30 (NC1 for (P_E)) *If x_* is a local minimum of (P_E) , if f and c are differentiable at x_* , and if c is qualified for representing X_E at x_* in the sense (1.53) below, then there exists a multiplier $\lambda_* \in \mathbb{F}$ such that*

$$\nabla_x \ell(x_*, \lambda_*) = 0, \tag{1.52a}$$

$$c(x_*) = 0. \tag{1.52b}$$

Definitions 1.31 1) The set X_E can be described by several functions c . Some are better than others. The constraint c is said to be *qualified* for representing X_E at $x \in X_E$ if

$$T_x X_E = T'_x X_E := \mathcal{N}(c'(x)). \tag{1.53}$$

There always holds $T_x X_E \subseteq T'_x X_E$, but equality is not necessarily guaranteed. Qualification holds if $c'(x)$ is surjective (hence, this is a sufficient condition of constraint qualification).

- 2) A vector λ_* in (1.52a) is called a *Lagrange multiplier* or *optimal multiplier* associated with x_* . The term “multiplier” comes from the fact that it multiplies the constraint in the Lagrangian.
- 3) A point x_* satisfying (1.52) for some $\lambda_* \in \mathbb{F}$ is said to be a *stationary point*. Sometimes, one says that the pair (x_*, λ_*) satisfying (1.52) is *stationary*. □

By definition, the set of optimal multipliers λ_* associated with a stationary point x_* of (P_E) is the set defined and denoted by

$$A_* := \{\lambda_* \in \mathbb{F} : \nabla f(x_*) + c'(x_*)^* \lambda_* = 0\}.$$

It is therefore an affine subspace of \mathbb{F} , which is nonempty by definition of the stationarity of x_* . Clearly, this set A_* is reduced to a singleton (uniqueness of the optimal multiplier associated with x_*) if and only if $c'(x_*)$ is surjective.

Proposition 1.32 (SC1 for a convex (P_E)) *Suppose that problem (P_E) is convex in the sense of the definition 1.29 and that f and c are differentiable at a point $x_* \in \mathbb{E}$ that satisfies (1.52) for some $\lambda_* \in \mathbb{F}$. Then, x_* is a global minimum of (P_E) .*

Second Order Optimality Conditions

First order necessary conditions allows us to select a set of points that are candidates for being solutions to (P_E) , but these conditions do not even discard a local maximum! With second order necessary conditions (NC2), the selection is more precise and many point that are not local minimums are removed from the list of stationary points.

Theorem 1.33 (NC2 for (P_E)) *If x_* is a local minimum of (P_E) , if f and c are twice differentiable at x_* , and if (1.52a) holds for some $\lambda_* \in \mathbb{F}$, then*

$$\forall d \in \mathbb{T}_{x_*} X_E : \langle \nabla_{xx}^2 \ell(x_*, \lambda_*) d, d \rangle \geq 0. \quad (1.54)$$

Remarks 1.34 1) With respect to the second order optimality conditions for the *unconstrained* problem “inf $f(x)$ ”, which read

$$\nabla f(x_*) = 0 \quad \text{and} \quad \nabla^2 f(x_*) \geq 0,$$

the NC2 for (P_E) have two major differences:

- it is the Lagrangian ℓ that intervenes in the conditions, not f ,
 - the Hessian of the Lagrangian is not positive semidefinite in the whole space \mathbb{E} but only in the tangent cone $\mathbb{T}_{x_*} X_E$, which is the subspace $\mathcal{N}(c'(x_*))$ if the constraint qualification (1.53) holds at x_* .
- 2) The result does not require any constraint qualification, but claims the inequality in (1.54) only for the tangent directions $d \in \mathbb{T}_{x_*} X_E$.
- 3) The inequality in (1.54) is not necessarily true for $d \in \mathcal{N}(c'(x_*)) \setminus \mathbb{T}_{x_*} X_E$ (but it holds in the presence of qualification, by the definition (1.53) of the latter). For example, if one minimizes $-x^2$ on $\{x \in \mathbb{R} : x^2 = 0\}$, the unique solution is the origin, $T_0 X_E = \{0\}$, $\mathcal{N}(c'(0)) = \mathbb{R}$, and $\nabla_x \ell(0, 0) = 0$, but $\nabla_{xx}^2 \ell(0, 0) = -2$ is positive negative on \mathbb{R} (while it is positive semidefinite on $\{0\}$).
- Of course, if the constraint c is qualified at x_* , one has $\mathbb{T}_{x_*} X_E = \mathcal{N}(c'(x_*))$ and the inequality in (1.54) is valid for all $d \in \mathcal{N}(c'(x_*))$.
- 4) As usual, these second order necessary conditions can be used to detect stationary points that are not local minimizers. \square

Sufficient optimality conditions of the second order (SC2) for the nonconvex problem (P_E) are essentially local; they do not guarantee global optimality (compare with proposition 1.32).

Theorem 1.35 (SC2 for (P_E)) *If f and c are twice differentiable at x_* , if (1.52) holds for some $\lambda_* \in \mathbb{F}$, and if*

$$\forall d \in \mathbb{T}_{x_*} X_E \setminus \{0\} : \quad d^\top \nabla_{xx}^2 \ell(x_*, \lambda_*) d > 0, \quad (1.55)$$

then x_ is a strict local minimum of (P_E) .*

Of course, (1.55) is stronger (hence the conclusion holds) if the inequalities hold for all $d \in \mathcal{N}(c'(x_*)) \setminus \{0\}$.

1.4.4 Equality and Inequality Constrained Problem (P_{EI})

Nonlinear optimization deals with the study of the *nonlinear optimization problem*, which consists in minimizing a function subject to equality and inequality constraints. A generic form of this problem is the following

$$(P_{EI}) \quad \begin{cases} \inf_x f(x) \\ c_E(x) = 0 \\ c_I(x) \leq 0, \end{cases}$$

where $f : \mathbb{E} \rightarrow \mathbb{R}$, $c_E : \mathbb{E} \rightarrow \mathbb{R}^{m_E}$, and $c_I : \mathbb{E} \rightarrow \mathbb{R}^{m_I}$ are (generally) smooth (possibly nonconvex) functions. Below it is considered that c_E and c_I are the components of a function $c : \mathbb{E} \rightarrow \mathbb{R}^m$; hence, E and I form a partition of $[1 : m]$, $m_E = |E|$, $m_I = |I|$, and $m_E + m_I = m$. As usual, the inequality “ $c_I(x) \leq 0$ ” has to be understood componentwise, meaning that it is equivalent to “ $c_i(x) \leq 0$ for all $i \in I$ ”. The *feasible set of (P_{EI})* is denoted by

$$X_{EI} := \{x \in \mathbb{E} : c_E(x) = 0, c_I(x) \leq 0\}.$$

Definition 1.36 The problem (P_{EI}) is said to be *convex* if f is convex and X_{EI} is a convex set. \square

Requiring the convexity of X_{EI} is less demanding than requiring the affinity of c_E and the componentwise convexity of c_I .

Proposition 1.37 (sufficient conditions for X_{EI} convex) *If c_E is affine and c_I is componentwise convex, then X_{EI} is convex.*

We say that the inequality constraint $c_i(x) \leq 0$, for some $i \in I$, is *active* at a point $x \in X_{EI}$ if $c_i(x) = 0$. The *sets of indices of active and inactive inequality constraints* are respectively denoted by

$$I^0(x) := \{i \in I : c_i(x) = 0\} \quad \text{and} \quad I^\sim(x) := \{i \in I : c_i(x) < 0\}.$$

We alleviate notation by setting $I_*^0 := I^0(x_*)$ and $I_*^\sim := I^\sim(x_*)$ for a given point $x_* \in X_{EI}$.

Constraint Qualification

The necessary condition of optimality of Peano-Kantorovich (1.51) makes use of the tangent cone to the feasible set. Our goal is to see how this condition reads in the case of the problem (P_{EI}) , so that an analytic expression of $T_x X_{EI}$ is desirable. As expected, this description of $T_x X_{EI}$ uses the derivatives of the function c defining the feasible set X_{EI} .

The tangent cone $T_x X_{EI}$ is always contained in the *linearizing cone*

$$T'_x X_{EI} := \{d \in \mathbb{E} : c'_E(x) \cdot d = 0, c'_{I^0(x)}(x) \cdot d \leq 0\}, \quad (1.56)$$

that is

$$T_x X_{EI} \subseteq T'_x X_{EI}.$$

One would like to have equality, since then an explicit expression of the tangent cone is at hand, which makes possible an analytic expression of the optimality (our goal in this section). This is not unusual, but not always true, since $T'_x X_{EI}$ is a *convex* polyhedron, while the tangent cone $T_x X_{EI}$ may not be convex.

Definition 1.38 (qualification of c to represent X_{EI}) One says that the constraint c is *qualified for representing X_{EI} at x* if

$$T_x X_{EI} = T'_x X_{EI}. \quad (1.57)$$

□

It is usually difficult to verify directly whether (1.57) holds. To make this frequent task easier, one has identify a number of *sufficient* conditions guaranteeing (1.57). Here are the four mostly used (there are many others).

- (CQ-A) [A for *Affinity*] $c_{E \cup I^0(x)}$ is affine near x ,
- (CQ-S) [S for *Slater* [130]] c_E is *affine*, $c_{I^0(x)}$ is componentwise convex, $\exists \check{x} \in X_{EI}$ such that $c_{I^0(x)}(\check{x}) < 0$,
- (CQ-LI) [LI for *Linear Independence*] $\sum_{i \in E \cup I^0(x)} \alpha_i \nabla c_i(x) = 0 \implies \alpha = 0$,
- (CQ-MF) [MF for *Mangasarian-Fromovitz* [97]] $\sum_{i \in E \cup I^0(x)} \alpha_i \nabla c_i(x) = 0$ and $\alpha_{I^0(x)} \geq 0 \implies \alpha = 0$.

It is clear that (CQ-LI) implies (CQ-MF), so that the latter is more often satisfied than the former.

Proposition 1.39 (other forms of (CQ-MF)) *Suppose that $c_{E \cup I^0(x)}$ is differentiable at $x \in X_{EI}$. Then, the following properties are equivalent:*

- (i) (CQ-MF) holds at x ,
- (ii) $\forall v \in \mathbb{R}^m, \exists d \in \mathbb{E}: c'_E(x) \cdot d = v_E$ and $c'_{I^0(x)}(x) \cdot d \leq v_{I^0(x)}$,
- (iii) $c'_E(x)$ is surjective and $\exists d \in \mathbb{E}: c'_E(x) \cdot d = 0$ and $c'_{I^0(x)}(x) \cdot d < 0$.

KKT Conditions

The following NC1 is often attributed to Karush, Kuhn, and Tucker [82, 87] (XXth century). It uses the *Lagrangian of (P_{EI})* , which is the function

$$\ell : (x, \lambda) \in \mathbb{E} \times \mathbb{F} \mapsto \ell(x, \lambda) = f(x) + \lambda^\top c(x) \in \mathbb{R}.$$

Because of its importance and because we shall follow the same track for determining NC1 for the general problem (P_G) in section 2.1.2, we provide a proof of the following result, which is very important in nonlinear optimization.

Theorem 1.40 (Karush-Kuhn-Tucker (KKT)) *If x_* is a local minimum of (P_{EI}) , if f and c are differentiable at x_* , and if c is qualified for representing X_{EI} at x_* in the sense (1.57), then, there exists a $\lambda_* \in \mathbb{R}^m$ such that*

$$\nabla_x \ell(x_*, \lambda_*) = 0, \quad (1.58a)$$

$$c_E(x_*) = 0, \quad (1.58b)$$

$$0 \leq (\lambda_*)_I \perp c_I(x_*) \leq 0. \quad (1.58c)$$

Proof. Let us alleviate notation by setting $J := I^0(x_*)$. We have successively

$$\begin{aligned} \nabla f(x_*) &\in (\mathbb{T}_{x_*} X_{EI})^+ && [(1.51)] \\ &= (\mathbb{T}'_{x_*} X_{EI})^+ && [\text{constraint qualification (1.57) at } x_*] \\ &= -\{c'_E(x_*)^* y + c'_J(x_*)^* z : y \in \mathbb{R}^{m_E}, z \in \mathbb{R}_+^{|J|}\} && [(1.2.8)]. \end{aligned}$$

Hence, there exist vectors $y \in \mathbb{R}^{m_E}$ and $z \in \mathbb{R}_+^{|J|}$ such that

$$\nabla f(x_*) = -c'_E(x_*)^* y - c'_J(x_*)^* z.$$

The conditions in (1.58) are then obtained by introducing

$$(\lambda_*)_i := \begin{cases} y_i & \text{if } i \in E \\ z_i & \text{if } i \in J \\ 0 & \text{if } i \in I \setminus J. \end{cases} \quad \square$$

Definitions 1.41 1) A vector λ_* appearing in (1.58) is called a *KKT* or *optimal multiplier* associated with x_* , since it multiplies the constraint in the Lagrangian.

The term “multiplier” comes from the fact that it multiplies the constraint in the Lagrangian.

- 2) A point x_* satisfying (1.58) is said to be a *stationary point*. Sometimes, one says that the pair (x_*, λ_*) satisfying (1.58) is *stationary*.
- 3) The condition (1.58c) is special and is known as the *complementarity conditions* of the problem (P_{EI}) . With their three operators, they mean the three conditions

$$(\lambda_*)_I \geq 0, \quad (\lambda_*)_I^\top c_I(x_*) = 0, \quad \text{and} \quad c_I(x_*) \leq 0.$$

Hence “ \perp ” expresses the orthogonality with respect to the Euclidean product in \mathbb{R}^{m_I} . Because of the sign of λ_* and $c_I(x_*)$, the second condition above has the following equivalent expressions

$$\begin{aligned} (\lambda_*)_I^\top c_I(x_*) = 0 &\iff \forall i \in I : (\lambda_*)_i c_i(x_*) = 0 \\ &\iff \forall i \in I : (c_i(x_*) < 0 \implies (\lambda_*)_i = 0). \end{aligned}$$

In other words, a multiplier associated with an inactive constraint vanishes.

4) One says that *strict complementarity* holds for a stationary pair (x_*, λ_*) if

$$\forall i \in I : (c_i(x_*) < 0 \iff (\lambda_*)_i = 0). \quad (1.59)$$

This is a property of (x_*, λ_*) , which is not necessary guaranteed by the fact that x_* is a local solution to (P_{EI}) . \square

By definition, the set of optimal multipliers λ_* associated with a stationary point x_* of (P_{EI}) is the set defined and denoted by

$$\Lambda_* := \{\lambda_* \in \mathbb{R}^m : \nabla f(x_*) + c'(x_*)^* \lambda_* = 0, (\lambda_*)_{I_*^0} \geq 0, (\lambda_*)_{I_*^{\sim}} = 0\}.$$

It is therefore a convex polyhedron of \mathbb{R}^m , which is nonempty by definition of the stationarity of x_* . The next proposition provides two properties of this set: a characterization of its boundedness [59] and of its uniqueness [89]. For a given multiplier λ_* , the following notation is adopted:

$$I_*^{0+} := \{i \in I_*^0 : (\lambda_*)_i > 0\} \quad \text{and} \quad I_*^{00} := \{i \in I_*^0 : (\lambda_*)_i = 0\}.$$

Proposition 1.42 (set of optimal multipliers) *Suppose that f and c are differentiable at a stationary point x_* of problem (P_{EI}) and that the set of associated multipliers Λ_* is nonempty. Then,*

- 1) Λ_* is bounded if and only if $(CQ-MF)$ holds,
- 2) for a given $\lambda_* \in \mathbb{R}^m$, the following properties are equivalent:

- (i) $\Lambda_* = \{\lambda_*\}$,
- (ii) any vector $\alpha \in \mathbb{R}^{|E|+|I_*^0|}$ verifying

$$\sum_{i \in E \cup I_*^0} \alpha_i \nabla c_i(x_*) = 0 \quad \text{and} \quad \alpha_{I_*^{00}} \geq 0$$

vanishes,

- (iii) $c'_{E \cup I_*^{0+}}(x_*)$ is surjective and there exists a vector $d \in \mathbb{E}$ such that

$$c'_{E \cup I_*^{0+}}(x_*)d = 0 \quad \text{and} \quad c'_{I_*^{00}}(x_*)d < 0.$$

The boundedness characterization is surprising, since Λ_* is linked to the optimization problem (P_{EI}) , while $(CQ-MF)$ only depends on the feasible set.

SC1 for a Convex Problem

Roughly speaking, the KKT conditions at a pair (x_*, λ_*) are sufficient to guarantee that x_* is a global minimum of (P_{EI}) , provided the problem is convex in the sense of the definition 1.36. Note that this CS1 does not require a constraint qualification.

Proposition 1.43 (SC1 for a convex (P_{EI})) Suppose that problem (P_{EI}) is convex in the sense of the definition 1.36 and that f and c are differentiable at a point $x_* \in \mathbb{E}$ that satisfies (1.58) for some $\lambda_* \in \mathbb{F}$. Then, x_* is a global minimum of (P_{EI}) .

1.4.5 Abstract Duality

This section describes various approaches allowing us to make links between two optimization problems, which may seem to have no connection to each other. Sometimes, the optimal multipliers of one problem are the solutions to the other one, and vice versa. The most often used scheme to connect two problems is the min-max duality.

Min-max duality

Consider the general optimization problem

$$(P) \quad \inf_{x \in X} f(x),$$

where X is an arbitrary set and $f : X \rightarrow \overline{\mathbb{R}}$ is an arbitrary function. Denote its optimal value and its solution set respectively by

$$\text{val}(P) \quad \text{and} \quad \text{Sol}(P).$$

In the duality context, (P) is called the *primal problem*, that is the problem that is first present.

Min-max duality comes into play when $f(x)$ can be written as a supremum:

$$f(x) = \sup_{y \in Y} \varphi(x, y),$$

where Y is some arbitrary set and $\varphi : X \times Y \rightarrow \overline{\mathbb{R}}$ is some arbitrary function. The function φ is called the *pairing function*. Then problem (P) reads

$$(P) \quad \inf_{x \in X} \sup_{y \in Y} \varphi(x, y).$$

The *dual problem* is then obtained by inverting the order in which the infimum and the supremum are taken:

$$(D) \quad \sup_{y \in Y} \inf_{x \in X} \varphi(x, y).$$

Denote its optimal value and its solution set respectively by

$$\text{val}(D) \quad \text{and} \quad \text{Sol}(D).$$

Aside from this audacious construction, which allowed us to derive the dual problem from the primal problem, the two problems may have no interesting links with each other. Nevertheless, the following so-called *weak duality inequality* always holds:

$$\text{val}(D) \leq \text{val}(P). \quad (1.60)$$

Proof. Clearly

$$\forall x' \in X, \forall y' \in Y : \quad \varphi(x', y') \leq \varphi(x', y')$$

and therefore certainly

$$\forall x' \in X, \forall y' \in Y : \quad \inf_{x \in X} \varphi(x, y') \leq \varphi(x', y').$$

Fixing $x' \in X$ and taking the supremum in $y' \in Y$ in the two sides, yields

$$\forall x' \in X : \quad \sup_{y \in Y} \inf_{x \in X} \varphi(x, y) \leq \sup_{y \in Y} \varphi(x', y).$$

Since the left-hand side is independent of x' , one can take the infimum in $x' \in X$ in the right-hand side and keep the inequality. This yields (1.60). \square

Given the very general context, this inequality is remarkable. When equality occurs in (1.60), one says that there is *no duality gap* between the primal and dual problems. This fact is not guaranteed. When the inequality (1.60) is strict, the positive value $\text{val}(P) - \text{val}(D)$ is called the *duality gap*; it may be infinite.

A stronger link occurs between the primal and dual problems when the pairing function φ has a *saddle-point*, which is a point $(\bar{x}, \bar{y}) \in X \times Y$ such that

$$\forall (x, y) \in X \times Y : \quad \varphi(\bar{x}, y) \leq \varphi(\bar{x}, \bar{y}) \leq \varphi(x, \bar{y}).$$

In other words, $\varphi(\bar{x}, \cdot)$ must have a maximum at \bar{y} and $\varphi(\cdot, \bar{y})$ must have a minimum at \bar{x} ; nothing is required outside the vertical cross $(\{\bar{x}\} \times Y) \cup (X \times \{\bar{y}\})$. A saddle-point may be characterized in terms of $\text{val}(P)$, $\text{Sol}(P)$, $\text{val}(D)$, and $\text{Sol}(D)$:

$$(\bar{x}, \bar{y}) \text{ is a saddle-point of } \varphi \iff \begin{cases} \bar{x} \in \text{Sol}(P) \\ \bar{y} \in \text{Sol}(D) \\ \text{val}(D) = \text{val}(P). \end{cases} \quad (1.61)$$

The set of saddle-points of the function φ is a Cartesian product, that is a set of the form $\bar{X} \times \bar{Y}$, where $\bar{X} \subseteq X$ and $\bar{Y} \subseteq Y$. This means that if (\bar{x}_1, \bar{y}_1) and (\bar{x}_2, \bar{y}_2) are saddle-points of φ , then (\bar{x}_1, \bar{y}_2) and (\bar{x}_2, \bar{y}_1) are also saddle-points of φ .

The primal and dual problems do not have the same solutions, since these live in different sets, in X for the former and in Y for the latter. Sometimes, the interest of the dual problem is that it may be easier to solve than the primal. Then, the question arises to know how to get a primal solution (what interests the designer of

that problem), when a solution to the dual problem is known. The next proposition explains how to get such a primal solution in the present very general context.

Proposition 1.44 *Suppose that φ has a saddle-point (\bar{x}, \bar{y}) . Then,*

$$\emptyset \neq \text{Sol}(P) \subseteq \underset{x \in X}{\text{Arg min}} \varphi(x, \bar{y}) \quad \text{and} \quad \emptyset \neq \text{Sol}(D) \subseteq \underset{y \in Y}{\text{Arg max}} \varphi(\bar{x}, y).$$

The first claim of the proposition tells us that, if φ has a saddle-point and if \bar{y} is a solution to the dual problem, the solutions to the primal problem are also the solutions to the problem $\inf\{\varphi(x, \bar{y}) : x \in X\}$. Therefore, there are some chance to recover a solution to the primal problem by solving the optimization problem $\inf\{\varphi(x, \bar{y}) : x \in X\}$. Now, this last problem may also have solutions that are not solution to (P) ; we call them *improper solutions*. Analog comments can be made for the second claim of the proposition.

1.4.6 Linear Optimization Problem (P_L)

Let \mathbb{E} be Euclidean vector space, $c \in \mathbb{E}$, $A : \mathbb{E} \rightarrow \mathbb{R}^m$ and $B : \mathbb{E} \rightarrow \mathbb{R}^p$ be linear maps, $a \in \mathbb{R}^m$, and $b \in \mathbb{R}^p$. A *linear optimization problem* (P_L) and its *Lagrangian dual* (D_L) read

$$(P_L) \quad \begin{cases} \inf_{x \in \mathbb{E}} \langle c, x \rangle \\ Ax = a \\ Bx \leq b \end{cases} \quad \text{and} \quad (D_L) \quad \begin{cases} \sup_{(y,s) \in \mathbb{R}^m \times \mathbb{R}^p} a^\top y - b^\top s \\ A^*y - B^*s = c \\ s \geq 0. \end{cases} \quad (1.62)$$

See below to learn how the dual problem is derived from the primal problem using min-max duality. The optimal value and the solution set of the primal problem are denoted by $\text{val}(P_L)$ and $\text{Sol}(P_L)$ respectively. Similarly, the optimal value and the solution set of the dual problem are denoted by $\text{val}(D_L)$ and $\text{Sol}(D_L)$, respectively.

The existence of a solution to the primal problem (P_L) is characterized by the following equivalence, in which $\text{val}(P_L) \in \mathbb{R}$ means that the problem (P_L) is feasible (i.e., its feasible set is nonempty, or, equivalently $\text{val}(P_L) < \infty$) and bounded (i.e., $\text{val}(P_L) > -\infty$).

$$(P_L) \text{ has a solution} \iff \text{val}(P_L) \in \mathbb{R}. \quad (1.63)$$

Proof. Since the left-to-right implication is clear, we only have to prove the reciprocal. Assume that $\text{val}(P_L) \in \mathbb{R}$. Then one can find a [minimizing sequence](#) (since the feasible set is nonempty), i.e., a sequence $\{x_k\}$ such that

$$Ax_k = a, \quad Bx_k \leq b, \quad \text{and} \quad \langle c, x_k \rangle \rightarrow \text{val}(P_L).$$

Furthermore, the set $\{\langle c, x \rangle : Ax = a, Bx \leq b\}$ is a closed interval in \mathbb{R} (since it is the image by the linear map $x \mapsto \langle c, x \rangle$ of the convex polyhedron $\{x \in \mathbb{E} : Ax = a, Bx \leq 0\}$, hence a convex polyhedron of \mathbb{R} ; see property 1 of proposition 1.1). Since

$\langle c, x_k \rangle$ is in that interval and converges to $\text{val}(P_L)$, this finite limit belongs to this interval, meaning that there exists a point $\bar{x} \in \mathbb{E}$ such that $\text{val}(P_L) = \langle c, \bar{x} \rangle$, $A\bar{x} = a$, and $B\bar{x} \leq b$. This point \bar{x} is therefore a solution to (P_L) . \square

The dual problem (D_L) is obtained from the primal problem (P_L) using the min-max duality, with the Lagrangian ℓ as *pairing function* (denoted φ in section 1.4.5). One can process that way since the primal problem can be written as the following infsup problem:

$$\inf_{x \in \mathbb{E}} \sup_{\substack{y \in \mathbb{R}^m \\ s \in \mathbb{R}_+^p}} \langle c, x \rangle - y^\top (Ax - a) + s^\top (Bx - b). \quad (1.64)$$

This is because

$$\sup_{\substack{y \in \mathbb{R}^m \\ s \in \mathbb{R}_+^p}} \langle c, x \rangle - y^\top (Ax - a) + s^\top (Bx - b) = \begin{cases} \langle c, x \rangle & \text{if } Ax = a \text{ and } Bx \leq b \\ +\infty & \text{otherwise.} \end{cases}$$

Indeed, first, if $Ax = a$ and $Bx \leq b$, $y^\top (Ax - a) = 0$ and $s^\top (Bx - b) \leq 0$ (since $s \geq 0$) and the supremum value $\langle c, x \rangle$ can be reached by taking any y and $s = 0$; next, if $Ax \neq a$, the infinite supremum is obtained by taking $y = t(Ax - a)$, $s = 0$, and $t \rightarrow \infty$; and finally, if $Bx \not\leq b$, the infinite supremum is obtained by taking $y = 0$, $s = t(Bx - b)^+ = t \max(0, Bx - b)$, and $t \rightarrow \infty$. Problem (D_L) is then obtained by inverting the infimum and the supremum in (1.64), to get

$$\begin{aligned} & \sup_{\substack{y \in \mathbb{R}^m \\ s \in \mathbb{R}_+^p}} \inf_{x \in \mathbb{E}} \langle c, x \rangle - y^\top (Ax - a) + s^\top (Bx - b) \\ &= \sup_{\substack{y \in \mathbb{R}^m \\ s \in \mathbb{R}_+^p}} \inf_{x \in \mathbb{E}} \langle c - A^*y + B^*s, x \rangle + a^\top y - b^\top s \\ &= \sup_{\substack{y \in \mathbb{R}^m \\ s \in \mathbb{R}_+^p}} \begin{cases} a^\top y - b^\top s & \text{if } A^*y - B^*s = c \\ -\infty & \text{otherwise} \end{cases} \\ &= \sup_{\substack{y \in \mathbb{R}^m \\ s \in \mathbb{R}_+^p \\ A^*y - B^*s = c}} a^\top y - b^\top s, \end{aligned}$$

which is indeed (D_L) .

A consequence of the above dualization process is the *weak duality* inequality (see (1.60)):

$$\boxed{\text{val}(D_L) \leq \text{val}(P_L)}. \quad (1.65)$$

The *strong duality* result refers to the following equivalences:

$$\boxed{(P_L) \text{ and } (D_L) \text{ are feasible} \iff \text{Sol}(P_L) \neq \emptyset \iff \text{Sol}(D_L) \neq \emptyset.}$$

(1.66)

When the conditions in (1.66) hold, the primal and dual optimal values are identical (there is no duality gap): $\text{val}(D_L) = \text{val}(P_L)$. The implication “ (P_L) and (D_L) are feasible $\Rightarrow \text{Sol}(P_L) \neq \emptyset$ ” is often used to show that the primal problem has a solution.

1.5 Algorithmics

1.5.1 Speeds of Convergence

Definitions 1.45 (speeds of convergence) Let \mathbb{E} be a normed space and $\{x_k\} \subseteq \mathbb{E}$ be a sequence converging to $x_* \in \mathbb{E}$, different from x_* . Then $\{x_k\}$ is said to converge

- *linearly*, if there exist a constant $r \in [0, 1)$ and an index $k_0 \in \mathbb{N}$ such that

$$\forall k \geq k_0 : \|x_{k+1} - x_*\| \leq r \|x_k - x_*\|,$$

- *superlinearly*, if

$$x_{k+1} - x_* = o(\|x_k - x_*\|),$$

- *quadratically*, if

$$x_{k+1} - x_* = O(\|x_k - x_*\|^2).$$

Remarks 1.46 1) The speeds of convergence above are sometimes named *quotient-speeds* since they are based on an estimation of the quotient $\|x_{k+1} - x_*\|/\|x_k - x_*\|$. Such an estimation is usually obtained by taking the development around x_* of the functions involved in the definition of the problem to solve and by using the algorithm definition and properties. Sometimes one uses the terms q-linear, q-superlinear, and q-quadratic convergence, to distinguish them from the r-linear, r-superlinear, and r-quadratic convergence, which are less demanding notions called *root-speeds* of convergence [106].

- 2) The property of linear convergence depends on the norm chosen on \mathbb{E} (linear convergence may occur for one norm and not for another one), but not those of superlinear and quadratic convergences.
- 3) Obviously, the superlinear convergence is faster than the linear convergence and the quadratic convergence is faster than the superlinear convergence.
- 4) Superlinear convergence is typically obtained by the quasi-Newton methods (when everything is going well), while quadratic convergence is typical of Newton's method.

To conclude this section, let us mention a property of superlinear convergence that will be useful in section 4.1.4. We say that two sequences $\{u_k\}_{k \geq 0}$ and $\{v_k\}_{k \geq 0}$ in a normed space \mathbb{E} *converge to zero equivalently*, a concept that we denote by $\{u_k\} \sim \{v_k\}$, if

$$\exists C > 0, \quad \forall k \geq 0 : \quad C^{-1} \|u_k\| \leq \|v_k\| \leq C \|u_k\|.$$

Lemma 1.47 (equivalent vanishing sequences) *If the sequence $\{x_k\}$ converges to x_* superlinearly, then, $\{x_{k+1} - x_k\} \sim \{x_k - x_*\}$.*

Proof. Just write $x_k - x_{k+1} = (x_k - x_*) - (x_{k+1} - x_*) = (x_k - x_*) + o(\|x_k - x_*\|)$ and conclude. \square

1.5.2 Newton and Quasi-Newton Algorithms ▲

Nonlinear Equation

Consider first the problem of finding a zero of a nonlinear system of equations. A function $F : \mathbb{E} \rightarrow \mathbb{F}$ between two vector spaces \mathbb{E} and \mathbb{F} of the same finite dimension is given and the problem consists in finding a *zero* of F , that is a point $x \in \mathbb{E}$ such that

$$F(x) = 0. \quad (1.67)$$

Newton's algorithm computes such a zero approximately (but with high precision) by generating a sequence $\{x_k\} \subseteq \mathbb{E}$, which is expected to converge to a zero of F . The procedure is as follows. Given the current iterate $x_k \in \mathbb{E}$, F is first linearized at x_k yielding the affine model of F :

$$x \in \mathbb{E} \mapsto F(x_k) + F'(x_k) \cdot (x - x_k),$$

which is a good approximation of F near x_k . Then, it makes sense to take as next iterate x_{k+1} , a zero of this affine map. Hence x_{k+1} solves (if possible)

$$F(x_k) + F'(x_k) \cdot (x_{k+1} - x_k) = 0. \quad (1.68a)$$

This is a linear system, which is much easier to solve than the original nonlinear system (1.67). If F is \mathcal{C}^1 and x_* is a *regular* zero, that is a zero with a nonsingular Jacobian $F'(x_*)$, then when the current iterate x_k is near x_* , $F'(x_k)$ is nonsingular and the next iterate x_{k+1} is obtained by

$$x_{k+1} = x_k - F'(x_k)^{-1} F(x_k). \quad (1.68b)$$

These claims are clarified by the local convergence result to theorem 1.48.

Conditions ensuring the local convergence of Newton's method are given in the next result.

Theorem 1.48 (convergence of the Newton algorithm) *Suppose that F has a zero x_* , that F is continuously differentiable around x_* , and that $F'(x_*)$ is nonsingular. Then*

- 1) *there exists $\varepsilon > 0$ such that, if the first iterate $x_1 \in \bar{B}(x_*, \varepsilon)$, the Newton algorithm (1.68), starting at x_1 , is well defined and generates a sequence $\{x_k\}$ converging superlinearly to x_* ,*
- 2) *if, furthermore, F is $\mathcal{C}^{1,1}$ in a neighborhood of x_* , the convergence is quadratic.*

A *quasi-Newton algorithm* operates like the Newton method, except that the Jacobians $F'(x_k)$ are approached by a linear operator $M_k : \mathbb{E} \rightarrow \mathbb{F}$, by a very specific technique. This avoids the need of computing the derivatives. Hence, the next iterate x_{k+1} is computed by solving the linear system

$$F(x_k) + M_k(x_{k+1} - x_k) = 0. \quad (1.69)$$

Despite the derivatives are not computed, superlinear convergence of the iterates is usually possible. To prove such speed of convergence, the following Dennis and Moré criterion for superlinear convergence is valuable. This criterion involves the quality of M_k only along the displacement direction $x_{k+1} - x_k$.

Proposition 1.49 (Dennis & Moré criterion for superlinear convergence) *Suppose F is differentiable at one of its zero x_* , that $F'(x_*)$ is nonsingular, and that $\{x_k\}$ generated by (1.69) converges to x_* . Then, the following properties are equivalent*

- (i) *the convergence of $\{x_k\}$ to x_* is superlinear,*
- (ii) $[M_k - F'(x_*)](x_{k+1} - x_k) = o(\|x_{k+1} - x_k\|)$.

Proof. Let us start by giving an expression of $[M_k - F'(x_*)](x_{k+1} - x_k)$ that takes into account the form of the algorithm displacement $(x_{k+1} - x_k)$, given by (1.69). There hold

$$\begin{aligned}
 & [M_k - F'(x_*)](x_{k+1} - x_k) \\
 &= -F(x_k) - F'(x_*)(x_{k+1} - x_k) \quad [(1.69)] \\
 &= -F'(x_*)(x_k - x_*) + o(\|x_k - x_*\|) - F'(x_*)(x_{k+1} - x_k) \\
 &\quad [F(x_*) = 0 \text{ and differentiability of } F \text{ at } x_*] \\
 &= -F'(x_*)(x_{k+1} - x_*) + o(\|x_k - x_*\|). \tag{1.70}
 \end{aligned}$$

[(i) \Rightarrow (ii)] By (i) and (1.70), we get

$$[M_k - F'(x_*)](x_{k+1} - x_k) = o(\|x_k - x_*\|).$$

Then (ii) follows by (i) and lemma 1.47.

[(ii) \Rightarrow (i)] If (ii) holds, (1.70) yield

$$F'(x_*)(x_{k+1} - x_*) = o(\|x_k - x_*\|).$$

The nonsingularity of $F'(x_*)$ then implies that $x_{k+1} - x_* = o(\|x_k - x_*\|)$, which is the superlinear convergence $\{x_k\}$, yielding (i). \square

Unconstrained Optimization \blacktriangle

Consider now the unconstrained optimization problem ...

1.5.3 Global Convergence in Unconstrained Optimization \blacktriangle

Consider the problem of minimizing a function $f : \mathbb{E} \rightarrow \mathbb{R}$ on a Euclidean vector space \mathbb{E} (no constraint):

$$\min_{x \in \mathbb{E}} f(x).$$

The *global convergence* of an algorithm generating a sequence of iterates $\{x_k\}_{k \geq 1}$ means that some kind of convergence result for this sequence $\{x_k\}_{k \geq 1}$ or most often

for some associated quantities (like the gradients $\nabla f(x_k)$) can be obtained, *whatever the initial iterate x_1 is*. This concept has to be compared with that of a *local convergence result*, which assumes that the initial iterate is close enough to a solution.

The most classical results “only” prove that the sequence of the gradients $\nabla f(x_k)$ converges to zero or, even less, that $\liminf_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0$ (meaning that a subsequence of $\{\nabla f(x_k)\}$ converges to zero). It is usually not the case that the sequence of iterates $\{x_k\}_{k \geq 1}$ itself converges to a solution to the problem, unless strong assumptions are taken on the function f , like its strong convexity. Nevertheless, if $\nabla f(x_k) \rightarrow 0$, any adherent point of $\{x_k\}_{k \geq 1}$ is a solution to the problem.

Once a method is able to define a direction of move $d_k \in \mathbb{E}$ at the current iterate x_k , a *globalization strategy* can be introduced, which is a technique able to force convergence from any starting point. The most famous globalization strategies are the use of *line-searches* and *trust-regions*.

Line-search

A *line-search algorithm* generates a sequence $\{x_k\}$ as follows:

- choice of a *descent direction* at x_k , i.e., $d_k \in \mathbb{E}$ such that $f'(x_k) \cdot d_k < 0$ (standard examples are gradient, conjugate gradient, Newton, quasi-Newton, and Gauss-Newton directions; each of these directions has specific properties and are more or less adapted to a given problem);
- determination of a *step-size* $\alpha_k > 0$ by a *line-search rule* to force a sufficient decrease of f along d_k (standard examples are Cauchy, Armijo, Goldstein, and Wolfe rules; each of these rules is adapted to particular situations, directions, and specificities of the problem at hand);
- the next iterate is then obtained by $x_{k+1} := x_k + \alpha_k d_k$.

This is a very simple algorithm, but with powerful properties. Its simplicity makes it more easily adaptable to problems that are more complex than unconstrained minimization.

In the next result, we simplify the notation by abbreviating $g_k := \nabla f(x_k)$.

Proposition 1.50 (global convergence with line-search) *Let $\{x_k\}$ be a sequence generated by the line-search algorithm described above. If $\{f(x_k)\}$ is bounded below, then*

$$\sum_{k \geq 1} \|g_k\|^2 \cos^2 \theta_k < +\infty, \quad (1.71)$$

where $\theta_k := \arccos \langle -g_k, d_k \rangle / (\|g_k\| \|d_k\|)$ is the angle between d_k and $-g_k$.

The property claiming that the series in (1.71) is convergent is called the *Zoutendijk condition* [138; 1970]. It provides the contribution of the line-search to the global convergence. The contribution of the directions d_k is also important and one cannot conclude any interesting properties without specifying it. Sometimes the analysis is long and tortuous [61; 1992]. The simplest situation is for the *gradient* or *steepest descent algorithm*, in which $d_k = -g_k$; in that case, $\cos \theta_k = 1$ and we obtain that the series $\sum_k \|g_k\|^2$ is convergent, implying in turn that $g_k \rightarrow 0$. This is the kind of results

that is highly desirable. Without assumption like the strong convexity of f , one cannot guarantee that the sequence of iterates $\{x_k\}$ converges to a point minimizing f (see [62; 2000] and the references therein). Hence, one can say that *the gradient algorithm with line-search is convergent*.

Trust-Region ▲

A *trust-region algorithm* generates a sequence $\{x_k\}$ as follows :

- choice of a *model* φ_k of f around x_k , usually quadratic: $\varphi_k(s) := \langle g_k, s \rangle + \frac{1}{2} \langle M_k s, s \rangle$ (standard models are the gradient, Newton, quasi-Newton, and Gauss-Newton models);
- determination of a trust-radius $\Delta_k > 0$ such that $f(x_k + s_k)$ is sufficiently smaller than $f(x_k)$, where $s_k \in \text{Arg min}\{\varphi_k(s) : \|s\| \leq \Delta_k\}$;
- the next iterate is then obtained by $x_{k+1} := x_k + s_k$.

Proposition 1.51 (global convergence with trust-region) *If f is \mathcal{C}^1 in a neighborhood of $V_1 := \{x \in \mathbb{E} : f(x) \leq f(x_1)\}$, if $\{f(x_k)\}$ is bounded below, and if $\{M_k\}$ is bounded, then $\liminf_{k \rightarrow \infty} \|g_k\| = 0$. If, furthermore, f is $\mathcal{C}^{1,1}$, then $g_k \rightarrow 0$.*

1.5.4 Global Convergence for Nonlinear Equations ▲

Let $F : \mathbb{E} \rightarrow \mathbb{E}$ and consider the problem of finding x such that

$$F(x) = 0.$$

No algorithm with global convergence (a few exceptions however; e.g., when F is polynomial, but the methods are expensive and restricted to problems with a rather small dimension).

$$\min_{x \in \mathbb{E}} \left(\varphi(x) := \frac{1}{2} \|F(x)\|_2^2 \right).$$

An *amazing but misleading fact* is the following. If $\nabla \varphi(x_k) \neq 0$, the Gauss-Newton direction

$$d_k^{\text{GN}} \in \text{Arg min}_{d \in \mathbb{E}} \|F(x_k) + F'(x_k)d\|_2^2$$

is a descent direction of φ at x_k . May yield false convergence. It is *better to use the trust-region approach*: $x_{k+1} = x_k + s_k$, where

$$s_k \in \text{Arg min}_{\|s\|_2 \leq \Delta_k} \|F(x_k) + F'(x_k)s\|_2^2,$$

with well adapted trust-radius $\Delta_k > 0$.

2 Optimality Conditions

There is no more general optimization problem than (P_X) in (1.47), in which an arbitrary function f is minimized on an arbitrary set X (well, there, the set was supposed to belong to a Euclidean space, which is already a major restriction). In this chapter, we consider a problem with a little more structure than (P_X) , but not much: an additional function c is used to represent the feasible set, which is now defined by “ $c(x) \in G$ ”, where c is an arbitrary function and G is a nonempty closed convex set. The goal of the analysis of this chapter is to clarify where and how this function c and the set G intervene in the optimality conditions. A remarkable outcome is that the theory presented in section 1.4.4 for the equality and inequality constrained problem (P_{EI}) can be nicely extended to this much more general problem. The abstraction brought by the considered model will allow us to better figure out the meaning of the optimality conditions, by highlighting their geometrical structure, and has the technical advantage of not forcing us to work with indices, which makes the proofs easier and more elegant. Furthermore, this generalization is also a means to make a few steps on the way towards the analysis of infinite dimensional problems, which are perfectly well defined with this abstract setting.

2.1 First Order Optimality Conditions for (P_G)

The first order optimality conditions of an optimization problem are frequently used. Since they express optimality by a mathematical system made of equalities, inequalities, inclusions, *etc*, which can be solved, they can be used to compute a solution analytically, i.e., on a piece of paper, to design algorithms to compute them, to conceive a stopping criterion for the latter, to have properties on the solutions of the considered problems. For these reasons, their setting is a very important step in the understanding of an optimization problem.

2.1.1 Definition of the General Problem

Let \mathbb{E} and \mathbb{F} be two Euclidean vector spaces. We consider the problem

$$(P_G) \quad \begin{cases} \min f(x) \\ c(x) \in G, \end{cases} \quad (2.1)$$

where $f : \mathbb{E} \rightarrow \mathbb{R}$ and $c : \mathbb{E} \rightarrow \mathbb{F}$ are smooth functions, and G is nonempty *closed convex* set in \mathbb{F} (not necessarily a cone). The letter c is used to recall that the function intervenes into the constraint and the letter G is introduced to avoid the more

appropriate but too frequent C and refers to the rather generality of this set. The *feasible set* of the problem is denoted by

$$X_G := \{x \in \mathbb{E} : c(x) \in G\} = c^{-1}(G).$$

Definition 2.1 (convex (P_G)) The optimization problem (P_G) is said to be *convex* if its objective f is convex and its feasible set X_G is convex.

The following implication holds for the multifunction $T : \mathbb{E} \multimap \mathbb{F} : x \mapsto c(x) - G$.

$$T \text{ is convex} \implies X_G = T^{-1}(0) \text{ is convex.} \quad (2.2)$$

Proof. Note the equivalences

$$x \in X_G \iff c(x) \in G \iff 0 \in T(x) \iff x \in T^{-1}(0).$$

Therefore $X_G = T^{-1}(0)$ and, according to (1.45), X_G is convex by the convexity of T^{-1} (implied by the one of T) and the convexity of the singleton $\{0\}$. \square

Examples 2.2 A number of optimization problems can be written in the form (P_G) . Let us mention a few.

- 1) The *nonlinear optimization problem (P_{EI})* , presented in section 1.4.4, can be written in the form (P_G) , by taking $\mathbb{F} = \mathbb{R}^m$ and $G = \{0_{\mathbb{R}^m E}\} \times \mathbb{R}_-^{m_I}$. Therefore, problem (P_G) generalizes problem (P_{EI}) .

As a problem a little more general than (P_G) , we have

$$(P_{E[l,u]}) \quad \begin{cases} \inf_x f(x) \\ c_E(x) = 0 \\ c_I(x) \in [l, u], \end{cases}$$

where, for some vectors l and $u \in \overline{\mathbb{R}}^{m_I}$, $[l, u] := \{v \in \mathbb{R}^{m_I} : l \leq v \leq u\}$. This problem reads like problem (P_G) , with $\mathbb{F} = \mathbb{R}^m$ and $G = \{0_{\mathbb{R}^m E}\} \times [l, u]$.

- 2) The *linear semidefinite optimization problem* reads

$$(P_{\text{SDO}}) \quad \begin{cases} \inf_{X \in \mathcal{S}^n} \langle C, X \rangle \\ A(X) = b \\ X \succeq 0, \end{cases}$$

where $C \in \mathcal{S}^n$, $A : \mathcal{S}^n \rightarrow \mathbb{R}^m$ is linear, $b \in \mathbb{R}^m$ and $X \succeq 0$ requires that $X \in \mathcal{S}_+^n$. This problem can be written in the form (P_G) by taking $\mathbb{E} = \mathcal{S}^n$, $\mathbb{F} = \mathbb{R}^m \times \mathcal{S}^n$, $c : X \in \mathcal{S}^n \rightarrow (A(X) - b, X) \in \mathbb{R}^m \times \mathcal{S}^n$ and $G = \{0_{\mathbb{R}^m}\} \times \mathcal{S}_+^n$.

The semidefinite optimization problem is analyzed in chapter 6.

- 3) Let \mathbb{E} and \mathbb{F} be two Euclidean vector spaces. A *composite optimization problem* is an optimization problem of the form

$$\min_{x \in \mathbb{E}} (g \circ F)(x), \quad (2.3)$$

where $F : \mathbb{E} \rightarrow \mathbb{F}$ is a differentiable function, $g : \mathbb{F} \rightarrow \overline{\mathbb{R}}$ is a function that may be nonsmooth and $(g \circ F)$ denotes the composition of g and F ; hence $(g \circ F)(x) = g(F(x))$.

This problem can also be written [110] $\inf_{(x,\alpha) \in \mathbb{E} \times \mathbb{R}} \{\alpha \in \mathbb{R} : g(F(x)) \leq \alpha\}$ or

$$\begin{cases} \inf_{(x,\alpha) \in \mathbb{E} \times \mathbb{R}} \alpha \\ (F(x), \alpha) \in \text{epi}(g), \end{cases} \quad (2.4)$$

so that it is of the form (P_G) in (2.1) with $\mathbb{E} \curvearrowright \tilde{\mathbb{E}} := \mathbb{E} \times \mathbb{R}$, $\mathbb{F} \curvearrowright \tilde{\mathbb{F}} := \mathbb{F} \times \mathbb{R}$, $c : (x, \alpha) \in \tilde{\mathbb{E}} \mapsto (F(x), \alpha) \in \tilde{\mathbb{F}}$ and $G = \text{epi}(g) \subseteq \tilde{\mathbb{F}}$. Note that the reformulation (2.4) of (2.3) in terms of (P_G) is valid without the need to have g smooth, since this function appears in (2.4) only through its epigraph. The set G will be nonempty closed convex set if $g \in \overline{\text{Conv}}(\mathbb{F})$.

Conversely, the problem (P_G) can also be written as a composite optimization problem [71, 110], since it reads

$$\inf_{x \in \mathbb{E}} f(x) + \mathcal{I}_G(c(x)),$$

where \mathcal{I}_G is the indicator function of the set G . This last optimization problem is of the form (2.3), with a function $F : x \in \mathbb{E} \mapsto (f(x), c(x)) \in \mathbb{R} \times \mathbb{F}$ and $g : (z, y) \in \mathbb{R} \times \mathbb{F} \mapsto z + \mathcal{I}_G(y) \in \overline{\mathbb{R}}$.

Therefore, problems (P_G) and (2.3) are equivalent.

See the subsection *Composite optimization* in section 2.1.6 for a short analysis of the composite optimization problem. \square

2.1.2 First Order Optimality Conditions

Necessary conditions of optimality of the first order (NC1) for problem (P_G) can be obtained by using the same approach as for problems with equality and inequality constraints, which starts with the Peano-Kantorovich condition (1.51) and culminates in the proof of theorem 1.40. There are some additional technical difficulties, however, but these can be overcome. This approach requires the computation of the tangent cone $T_x X_G$ to the feasible set X_G at x . The first step consists in establishing a link between this tangent cone and the linearizing cone to X_G at x , which is the cone defined and denoted by

$$T'_x X_G := \{d \in \mathbb{E} : c'(x)d \in T_{c(x)} G\}.$$

Since it can also be written $c'(x)^{-1}[T_{c(x)} G]$, the linearizing cone is a nonempty closed convex cone, when G is a nonempty closed convex set.

Proposition 2.3 (tangent and linearizing cones) *If c is differentiable at $x \in X_G$, then*

$$T_x X_G \subseteq T'_x X_G. \quad (2.5)$$

Proof. Let $d \in \mathbb{T}_x X_G$. One may assume that $d \neq 0$, since $0 \in \mathbb{T}'_x X_G$, trivially. Then, there exist a sequence $\{x_k\} \subseteq X_G$ converging to x and a sequence $\{t_k\} \downarrow 0$ such that $(x_k - x)/t_k \rightarrow d$. Furthermore, c being differentiable at x , one can write $c(x_k) = c(x) + c'(x)(x_k - x) + o(\|x_k - x\|)$, so that

$$\frac{c(x_k) - c(x)}{t_k} \rightarrow c'(x)d.$$

Since $c(x) \in G$ and $c(x_k) \in G$, we deduce from this limit that $c'(x)d \in \mathbb{T}_{c(x)} G$. \square

Equality does not necessarily hold in (2.5), since $\mathbb{T}'_x X_G$ is a convex set (see the remark before the proposition), while $\mathbb{T}_x X_G$ is not necessarily convex (we have not required the affinity of the function c defining X_G). This is annoying, since it is the tangent cone $\mathbb{T}_x X_G$ that intervenes in the generic first order necessary condition of Peano-Kantorovich (1.51), while one would like to take advantage of the analytic expression of the linearizing cone $\mathbb{T}'_x X_G$. Like for the problem (PEI) (see the subsection *Constraint Qualification* in section 1.4.4), the notion of qualification of the function c to represent X_G is linked to the fact that equality holds in (2.5), *but it is not limited to that property*. Recall indeed the technique of proof of proposition 1.40 yielding the first order optimality conditions of Karush, Kuhn, and Tucker, a technique that will be also used to prove proposition 2.6 below. The goal is to show that the gradient $\nabla f(x_*)$ belongs to a cone that we want to be explicit. Two ingredients intervene in this approach:

- the equality between the tangent and linearizing cones, which allows us to make good use of the expression of the tangent cone given by the linearizing cone,
- the polyhedrality of the linearizing cone, which allows us to discard the closure operator acting on the set resulting from the use of the Farkas identity (1.15).

In the present case, $\mathbb{T}'_x X_G$ is not necessary polyhedral, because we do not want to impose this restrictive polyhedrality property to G . In order to select the nonconvex feasible sets for which the proposed approach to establish the first order optimality conditions can be used, we introduce a so-called *constraint qualification assumption*, which precisely ensures the equality between the tangent and linearizing cones (it is (2.6a) below, nothing new there with respect to the definition 1.38), but also the closed character of the image by $c'(x)^*$ of the dual of the tangent cone $\mathbb{T}_{c(x)} G$ (it is (2.6b) below).

Definition 2.4 (qualification of c to represent X_G) The constraint c is said to be *qualified for representing X_G at $x \in X_G$* if c is differentiable at x and if the following two conditions are satisfied:

$$\mathbb{T}_x X_G = \mathbb{T}'_x X_G, \tag{2.6a}$$

$$c'(x)^*[(\mathbb{T}_{c(x)} G)^+] \text{ is closed,} \tag{2.6b}$$

where $c'(x)^* : \mathbb{F} \rightarrow \mathbb{E}$ denotes the **adjoint** linear operator of $c'(x) : \mathbb{E} \rightarrow \mathbb{F}$. \square

The verification of this constraint qualification (2.6) is a difficult task. In order to simplify it, a *sufficient* condition of qualification is introduced and studied in sections 2.1.3 and 2.1.4.

Constraint qualifications make possible, like for problem (P_E) and (P_{EI}) , the obtention of first order optimality conditions, those of theorem 2.6 below. When $G = K$ is un cone, we shall use the complementarity notation

$$K^+ \ni v \perp u \in K$$

to mean, in a compact manner, that the vectors u and v of \mathbb{F} verify the three properties $u \in K$, $v \in K^+$ and $\langle u, v \rangle = 0$. We shall use the following equivalence.

Lemma 2.5 *Let K be a closed convex cone in a Euclidean space \mathbb{E} . Then*

$$v + N_u K \ni 0 \quad \Longleftrightarrow \quad K^+ \ni v \perp u \in K. \quad (2.7)$$

Proof. Since $N_u K = \emptyset$ if $u \notin K$, the condition $v + N_u K \ni 0$ is equivalent to

$$u \in K \quad \text{and} \quad \left(\langle -v, u' - u \rangle \leq 0, \quad \forall u' \in K \right).$$

Since K is a cone and $u \in K$, one can take $u' = 2u$ and $u' = \frac{1}{2}u$ (or $u' = 0$, since 0 belongs to the *closed* cone K) in the last condition, which therefore implies that $\langle u, v \rangle = 0$. As a result, the previous conditions are equivalent to

$$u \in K, \quad \langle u, v \rangle = 0, \quad \text{and} \quad \left(\langle v, u' \rangle \geq 0, \quad \forall u' \in K \right).$$

It remains to observe that the last condition also reads $v \in K^+$. □

Theorem 2.6 (NC1 for problem (P_G)) *Let x_* be a local solution to problem (P_G) . Suppose that f and c are differentiable at x_* and that c is qualified to represent X_G at x_* , in the sense of definition 2.4. Then, there exist $\lambda_* \in \mathbb{F}$ such that*

$$\nabla f(x_*) + c'(x_*)^* \lambda_* = 0, \quad (2.8a)$$

$$\lambda_* \in N_{c(x_*)} G. \quad (2.8b)$$

If $G \equiv K$ is also a cone, (2.8b) becomes

$$K^- \ni \lambda_* \perp c(x_*) \in K. \quad (2.8c)$$

Proof. Structurally, the proof is similar to the one of theorem 1.40, in a more abstract setting, however. We have successively

$$\begin{aligned} \nabla f(x_*) &\in (T_{x_*} X_G)^+ && \text{[Peano-Kantorovich (1.51)]} \\ &= (T'_{x_*} X_G)^+ && \text{[constraint qualification (2.6a) at } x_*] \\ &= \{d \in \mathbb{E} : c'(x_*)d \in T_{c(x_*)} G\}^+ && \text{[formula (2.5) of } T'_{x_*} X_G]. \end{aligned}$$

We now apply the Farkas identity (1.15) with $A^* = c'(x_*)$ and $K^+ = T_{c(x_*)} G$. Since the latter is a closed convex cone, the following holds $K = K^{++}$ (see (1.16)). We get

$$\nabla f(x_*) \in \overline{c'(x_*)^* [(T_{c(x_*)} G)^+]} = c'(x_*)^* [(T_{c(x_*)} G)^+],$$

where we have used the constraint qualification (2.6b) at $x = x_*$. We have shown the existence of a $\lambda_* \in -(T_{c(x_*)} G)^+ = N_{c(x_*)} G$ (it is (2.8b)) such that (2.8a) holds.

The complementarity condition (2.8c) results from (2.8b) and the equivalence (2.7), which can be applied since $G = K$ is a convex cone. \square

Definitions 2.7 (stationary point) A point x_* satisfying (2.8a)-(2.8b) for some $\lambda_* \in \mathbb{F}$ is said to be a *stationary point* of problem (P_G) . Sometimes, one says that the pair (x_*, λ_*) satisfying (2.8a)-(2.8b) is *stationary* for problem (P_G) . The vectors λ_* in (2.8a)-(2.8b) are called the *optimal multipliers* associated with x_* . \square

Remarks 2.8 1) If $c : \mathbb{E} \rightarrow \mathbb{E}$ is the identity, (2.8a)-(2.8b) reads $\nabla f(x_*) \in -N_{x_*} G = (T_{x_*} G)^+$ and one recovers the first order necessary condition of optimality of Peano-Kantorovich (1.51). In other words, the conditions (2.8a)-(2.8b) offer a way of taking into account non simple constraints, in which a function c intervenes.

2) One recognizes in (2.8a) the gradient with respect to x of the *Lagrangian of problem* (P_G) , which is the function $\ell : \mathbb{E} \times \mathbb{F} \rightarrow \mathbb{R}$ defined at $(x, \lambda) \in \mathbb{E} \times \mathbb{F}$ by

$$\ell(x, \lambda) = f(x) + \langle \lambda, c(x) \rangle. \quad (2.9)$$

This Lagrangian makes use of the functional part of the constraint $c(x) \in G$. The set G is taken into account by the second condition (2.8b).

3) The condition (2.8c) is called the *complementarity condition* of problem (P_G) , which therefore holds when $G \equiv K$ is a nonempty closed convex cone.

4) Problem (P_{EI}) can be written like problem (P_G) with the *polyhedral cone* $G \equiv K = \{0_{\mathbb{R}^{m_E}}\} \times \mathbb{R}_-^{m_I}$. Then, the optimality conditions have the form (2.8a) and (2.8c). Since $K^- = \mathbb{R}^{m_E} \times \mathbb{R}_+^{m_I}$, the complementarity condition (2.8c) only intervenes on the inequality constraints and reads $0 \leq (\lambda_*)_I \perp c_I(x_*) \leq 0$. One recovers the complementarity conditions of the KKT system (1.58). \square

A sibling of the sufficient optimality condition of the first order (SC1) of proposition 1.43 is given below. Like before, this SC1 does not require a constraint qualification. Recall also that the required convexity of X_G is ensured by the convexity of the multifunction $x \mapsto c(x) - G$, see (2.2).

Proposition 2.9 (SC1 for a convex (P_G)) Suppose that problem (P_G) is convex in the sense of the definition 2.1, that f and c are differentiable at $x_* \in X_G$, and that there exists a multiplier $\lambda_* \in \mathbb{F}$ such that (x_*, λ_*) verifies the first order optimality conditions (2.8a)-(2.8b). Then, x_* is a global solution to (P_G) .

Proof. By proposition 1.28, since f is a convex function and X_G is a convex set, it suffices to show that

$$\forall x \in X_G : \quad \langle \nabla f(x_*), x - x_* \rangle \geq 0.$$

By the first optimality condition (2.8a), this amounts to show that

$$\forall x \in X_G : \quad \langle \lambda_*, c'(x_*)(x - x_*) \rangle \leq 0.$$

By the second optimality condition (2.8b), $\lambda_* \in N_{c(x_*)} G = (T_{c(x_*)} G)^-$, so that it suffices to show that

$$\forall x \in X_G : \quad c'(x_*)(x - x_*) \in T_{c(x_*)} G.$$

Now, by definition of the derivative

$$c'(x_*)(x - x_*) = \lim_{t \downarrow 0} \frac{1}{t} [c(x_* + t(x - x_*)) - c(x_*)].$$

For $t \in [0, 1]$, $c(x_* + t(x - x_*)) = c((1-t)x_* + tx) \in G$ (since x_* and x are in the convex set X_G), so that the right-hand side is in the closure of $\mathbb{R}_+(G - c(x_*))$, which is $T_{c(x_*)} G$. \square

2.1.3 Robinson's Condition

It is not easy to verify that a given function c is qualified to represent X_G at $x \in X_G$, in the sense of definition 2.4. Like for the classical problems (P_E) and (P_{EI}) , one knows sufficient conditions ensuring that qualification. The most famous of them (there are reasons for that) is the *Robinson condition* [119], which at $x_0 \in X_G$ reads

$$(CQ-R) \quad 0 \in \text{int}(c(x_0) + c'(x_0)\mathbb{E} - G). \quad (2.10)$$

The goal of this section is to analyze and clarify this very rich condition. To arouse the curiosity of the reader, let us mention that (2.10) reduces to the Mangasarian-Fromovitz condition (CQ-MF) when the problem (P_{EI}) is viewed as a particular instance of problem (P_G) ; see exercise 2.1.4.

The condition (CQ-R) may look abstruse at first glance, but it is worth the effort to get familiar with it, since it is useful on several accounts. Its clarification and its application to the constraint qualification lie on the following important equivalence, which makes a link between (CQ-R) and the notion of metric regularity of the multifunction $x \in \mathbb{E} \mapsto c(x) - G \subseteq \mathbb{F}$, associated with the constraint of (P_G) . Recall the definition and notation (1.1) of the distance from a point to a set.

Theorem 2.10 (Robinson qualification and metric regularity) *If c is continuously differentiable in a neighborhood of $x_0 \in X_G$, then the following properties are equivalent:*

- (i) *the Robinson condition (CQ-R) holds at x_0 ,*
- (ii) *there exists a constant $\mu \geq 0$, such that $\forall (x, y) \in \mathbb{E} \times \mathbb{F}$ near $(x_0, 0)$, the following inequality holds*

$$\text{dist}(x, c^{-1}(G + y)) \leq \mu \text{dist}(c(x), G + y). \quad (2.11)$$

The meaning of the metric regularity property, given in the condition (ii) of the preceding theorem, is illustrated in figure 2.1: $G + y$ is a small perturbation (translation)

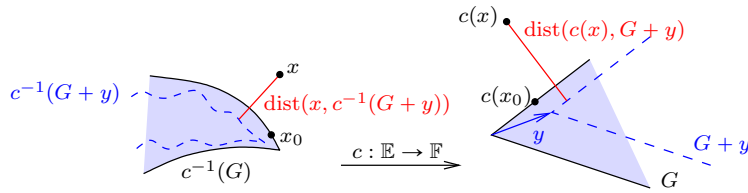


Fig. 2.1. Illustration of the notion of metric regularity for the set X_G .

of the closed convex set G , $c^{-1}(G + y)$ is then a small perturbation of the possibly nonconvex feasible set $c^{-1}(G)$, and the distance from x to the latter is estimated by means of the distance from $c(x)$ to the former. Practically, this estimate may be interesting, since $\text{dist}(c(x), G + y)$ (a distance to a convex set) is usually more easily computed than $\text{dist}(x, c^{-1}(G + y))$ (a distance to a possibly weird set).

Before giving a proof of this important theorem, which is long and complex, we start by giving two of its straightforward corollaries. This is a way of getting familiar with its meaning.

The first corollary gives an error bound for X_G . An *error bound* is an estimate of the distance to a set by a quantity more easily computable (numerically or analytically). In general, X_G is a set that is more “complex” than G , if only because G is convex, which may not be the case of $X_G = c^{-1}(G)$, which can be tortuous through the action of the arbitrary function c . The error bound of Robinson (2.12) below gives an estimate of the distance $\text{dist}(x, X_G)$, which may be difficult to compute, by the much simpler distance $\text{dist}(c(x), G)$.

Corollary 2.11 (Robinson’s error bound for X_G) *If c is continuously differentiable in a neighborhood of $x_0 \in X_G$ and if Robinson’s constraint qualification (CQ-R) holds at x_0 , then, there exists a constant $\mu \geq 0$, such that, for all x*

near x_0 , the following inequality holds

$$\text{dist}(x, X_G) \leq \mu \text{dist}(c(x), G). \tag{2.12}$$

Proof. Just take $y = 0$ in (2.11) and use the fact that $X_G = c^{-1}(G)$. □

Robinson’s error bound (2.12) will play a major part in the proof of the fact that (CQ-R) is a sufficient condition of constraint qualification (proposition 2.21).

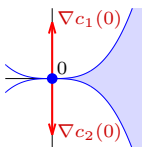
The inequalities (2.11) of theorem 2.10 form a family of error bounds for the sets $c^{-1}(G + y)$, with y small enough, which are small perturbations of X_G . Theorem 2.10 claims that this family of error bounds is an equivalent condition to (CQ-R), while its corollary 2.11 tells us that the single error bound (2.12) is a consequence of (CQ-R).

The distance to an empty set is infinite (see the remark after (1.1)). Since the distance in the right-hand side of (2.11) is finite (G is nonempty), the one in the left-hand side is also finite. This means that $c^{-1}(G + y)$ is nonempty for small perturbations y of G . It is this stability result for small perturbations that is claimed in the second corollary. The comments in this paragraph serve as its proof.

Corollary 2.12 (stability of X_G for small perturbations) *If c is continuously differentiable in a neighborhood of $x_0 \in X_G$ and if Robinson’s constraint qualification (CQ-R) holds at x_0 , then, for any perturbation $y \in \mathbb{F}$ close to 0, the following holds*

$$\{x \in \mathbb{E} : c(x) \in G + y\} \neq \emptyset. \tag{2.13}$$

Remark that (2.13) makes no reference to the point x_0 , which can therefore be an arbitrary point in X_G satisfying (CQ-R). Exercise 2.1.4 tells us that, for (P_{E1}) , (CQ-R) is equivalent to (CQ-MF), so that the following example shows that the reciprocal of the claim of corollary 2.12 is false: stability does not imply (CQ-R). Consider indeed the following set

$$\{x \in \mathbb{R}^2 : -x_1^3 \leq x_2 \leq x_1^3\}.$$


This one reads $\{x \in \mathbb{R}^2 : c(x) \in G\}$ with $c : \mathbb{R}^2 \rightarrow \mathbb{R}^2 : (x_1, x_2) \mapsto (-x_1^3 + x_2, -x_1^3 - x_2)$ and $G = \mathbb{R}_-^2$. It is nonempty, whatever the translation y of G is, while (CQ-MF) does not hold at $x = 0$ (because the gradients of the constraints $\nabla c_1(0)$ and $\nabla c_2(0)$ are collinear and opposite).

The rest of this section is entirely dedicated to the proof of theorem 2.10, whose generality suggests its strength, but also augurs the difficulty of its analysis. The hurried reader or the reader not interested in the proof can go directly to the following section 2.1.4. This one only uses from this section the corollary 2.11.

The proof of theorem 2.10, which will be given on page 70, is the consequence of two interpretations and two main results (theorem 2.15 and proposition 2.19). The analysis uses the multifunction toolbox. The following two multifunctions will play a major part:

$$T : \mathbb{E} \multimap \mathbb{F} : x \mapsto T(x) := c(x) - G, \quad (2.14a)$$

$$T_0 : \mathbb{E} \multimap \mathbb{F} : x \mapsto T_0(x) := c(x_0) + c'(x_0) \cdot (x - x_0) - G. \quad (2.14b)$$

The multifunction T was already used in (2.2) to provide a sufficient condition ensuring the convexity of the feasible set $X_G = T^{-1}(0)$, while T_0 is a kind of linearization of T at x_0 , in which only c is linearized, not G . One can now give the scheme of the proof of theorem 2.10.

1. *First interpretation.* Since the range of T_0 is $c(x_0) + c'(x_0)\mathbb{E} - G$, Robinson's condition (CQ-R) at x_0 reads

$$0 \in \text{int } \mathcal{R}(T_0).$$

This justifies the introduction of T_0 .

2. *First result.* For a *convex multifunction*, like T_0 (but not $T!$), this last condition turns out to be equivalent to the *metric regularity* of T_0 at $(x_0, 0) \in \mathcal{G}(T_0)$ (theorem 2.15). This means that there exists a constant $\mu_0 \geq 0$ such that, for all (x, y) near $(x_0, 0)$, the following inequality holds

$$\text{dist}(x, T_0^{-1}(y)) \leq \mu_0 \text{dist}(y, T_0(x)).$$

This notion is clarified below, after its definition 2.14.

3. *Second result.* When x is close to x_0 , T is “close” to T_0 near x_0 (and reciprocally), in a sense that will be made precise below. Then, one uses the fact that the metric regularity of T_0 is “diffused” to T (proposition 2.19): there exists a constant $\mu \geq 0$ such that, for (x, y) near $(x_0, 0)$, the following inequality holds

$$\text{dist}(x, T^{-1}(y)) \leq \mu \text{dist}(y, T(x)).$$

4. *Second interpretation.* To get (2.11), it suffices now to observe that

$$T^{-1}(y) = c^{-1}(G + y) \quad \text{and} \quad \text{dist}(y, T(x)) = \text{dist}(c(x), G + y).$$

The first identity results from the equivalences $x' \in T^{-1}(y) \Leftrightarrow y \in T(x') = c(x') - G \Leftrightarrow c(x') \in G + y \Leftrightarrow x' \in c^{-1}(G + y)$; while the second identity comes from $\text{dist}(y, T(x)) = \text{dist}(y, c(x) - G) = \text{dist}(c(x), G + y)$.

This proof scheme will be resumed in the proof of theorem 2.10 at the end of the section (on page 70). Let us now establish the results that will allow us to implement it.

We start by giving expressions that are equivalent to the fact that a point y_0 is interior to the range of a *convex multifunction*. As explained in the proof scheme aforementioned (see its steps 1 and 2), this result will be applied to the multifunction T_0 defined in (2.14b), not to T defined in (2.14a), which is generally nonconvex. We shall need the two important multifunction properties described by the definitions 2.13 and 2.14 below.

Definition 2.13 (open multifunction) A multifunction $T : \mathbb{E} \multimap \mathbb{F}$ is said to be *open* at $(x_0, y_0) \in \mathcal{G}(T)$ with ratio $\rho > 0$, if there exists a neighborhood W of (x_0, y_0) and a maximal radius $r_{\max} > 0$, such that, for all $(x, y) \in W \cap \mathcal{G}(T)$ and all $r \in [0, r_{\max}]$, the following holds

$$y + \rho r \bar{B}_{\mathbb{F}} \subseteq T(x + r \bar{B}_{\mathbb{E}}). \tag{2.15}$$

The ratio ρ is the quotient between the radius of the ball $y + \rho r \bar{B}_{\mathbb{F}}$ of \mathbb{F} that can be inscribed in the image $T(x + r \bar{B}_{\mathbb{E}})$ and the radius r . \square

By taking $r = 0$ in the inclusion (2.15), it appears that (x, y) must necessarily belong to the graph of T , and this fact is indeed assumed in the definition.

The picture in the left-hand side of figure 2.2 illustrates the notion of *open mul-*

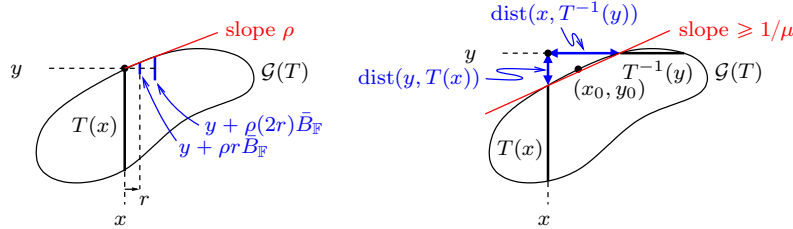


Fig. 2.2. Illustration of the notions of open multifunction of ratio ρ (left) and of metric regularity of modulus μ (right): both notions describe the variation of $T(x)$ with x ; the openness property makes this description from inside the graph, the metric regularity property makes it from outside; these are not infinitesimal concepts.

tifunction. One can see the balls centered at y in $T(x + r \bar{B}_{\mathbb{E}})$ and $T(x + 2r \bar{B}_{\mathbb{E}})$, for a certain radius $r > 0$ (they are translated for giving more visibility). The increase rate of the radius of these balls with r provides the ratio ρ . This ratio is not an infinitesimal notion. Thus, it does not provide the infinitesimal variation of the size of $T(x + r \bar{B}_{\mathbb{E}})$ with r , at $r = 0$, but provides an approximation of this variation from inside the graph of T (the pairs (x, y) belong to this graph).

Whilst the openness of T describes the variation of $T(x)$ with x , from *inside* the graph $\mathcal{G}(T)$, the metric regularity defined below provides a similar description, but from *outside* that graph.

Definition 2.14 (metric regular multifunction) A multifunction $T : \mathbb{E} \multimap \mathbb{F}$ is said to be *metric regular* at $(x_0, y_0) \in \mathcal{G}(T)$ of modulus $\mu > 0$, if for all (x, y) near (x_0, y_0) , the following inequality holds

$$\text{dist}(x, T^{-1}(y)) \leq \mu \text{dist}(y, T(x)). \tag{2.16}$$

This inequality assumes that $\text{dist}(\cdot, \emptyset) = +\infty$. \square

The picture in the right-hand side of figure 2.2 illustrates the notion of *metric regularity* and its link with the notion of open multifunction. Note first that, to have an estimate (2.16) bringing information, it is necessary to have (x, y) outside the

graph $\mathcal{G}(T)$ (otherwise $y \in T(x)$, $y \in T^{-1}(y)$, and the two distances in (2.16) vanish), hence not to have (x_0, y_0) in the interior of that graph, but on its boundary, like in the picture. We see that the modulus $\text{dist}(y, T(x))/\text{dist}(x, T^{-1}(y)) \simeq 1/\mu$ has a meaning that is similar to the ratio ρ in the left-hand side picture.

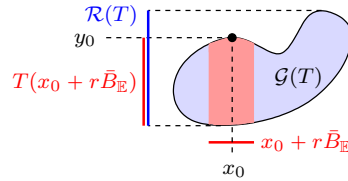
Note that (2.15) and (2.16) contain existence results. Indeed, when (2.15) holds and $y' \in y + \rho r \bar{B}_{\mathbb{F}}$, then, there exists an $x' \in x + r \bar{B}_{\mathbb{E}}$ such that $y' \in T(x')$: hence, the inclusion problem “given y' , find an x' such that $y' \in T(x')$ ” has a solution. Similarly, when (2.16) holds, $\text{dist}(x, T^{-1}(y))$ is finite (because $\text{dist}(y, T(x))$ is always finite), which means that $T^{-1}(y) \neq \emptyset$ or that there exists an x' such that $y \in T(x')$. In the proof of the theorem below, we use these openness and metric regularity properties in that manner.

Theorem 2.15 (open multifunction) *Let $\bar{B}_{\mathbb{E}}$ and $\bar{B}_{\mathbb{F}}$ be the closed unit balls of \mathbb{E} and \mathbb{F} respectively, $T : \mathbb{E} \multimap \mathbb{F}$ be a convex multifunction, and $(x_0, y_0) \in \mathcal{G}(T)$. Then, the following properties are equivalent:*

- (i) $y_0 \in \text{int } \mathcal{R}(T)$,
 - (ii) for any $r > 0$, one has $y_0 \in \text{int } T(x_0 + r \bar{B}_{\mathbb{E}})$,
 - (iii) T is open at (x_0, y_0) with some ratio $\rho > 0$,
 - (iv) T is metric regular at (x_0, y_0) with some modulus $\mu > 0$.
- One can take $\mu = 1/\rho$ in point (iv) if ρ is given by point (iii).

Before proving this theorem, let us make some remarks on its assumptions and its meaning.

- The open multifunction theorem above generalizes to convex multifunctions, which is a nonlinear object, the open mapping theorem on linear continuous maps, which is a basic tool in functional analysis [24]. This latter result is proposed in the exercise 2.1.2 as an equivalence between three claims (i)-(iii) (its proof in finite dimension is easy). Its claim (i) is comparable to point (i) above; its point (ii) is similar to point (ii) and (iii) above and its point (iii) is related to point (iv) above.
- The convexity of T cannot be discarded without losing the implications (i) \Rightarrow (ii) and (i) \Rightarrow (iii). Indeed, for the nonconvex multifunction T , whose graph is given in the figure below, $y_0 \in \text{int } \mathcal{R}(T)$, but $y_0 \notin \text{int } T(x_0 + r \bar{B}_{\mathbb{E}})$ for small $r > 0$;



furthermore, T is not open at (x_0, y_0) , since $T(x_0 + r \bar{B}_{\mathbb{E}})$ does not contain any ball in \mathbb{F} centered at y_0 .

- By its equivalence (i) \Leftrightarrow (iv), the theorem allows us to translate (CQ-R) in a metric regularity property, which is what we wanted to do, but the equivalence is limited to convex multifunctions for the while.

- The equivalence between points (iii) and (iv) makes explicit the link between open multifunction and metric regularity, in a rigorous manner. Thus, a large value of ρ testifies a fast variation of $T(x)$ with x , while a large value of μ expresses a slow variation of $T(x)$ with x .

Proof of proposition 2.15. [(i) \Rightarrow (ii)] Let $r > 0$. By the convexity de T and (1.45), $T(x_0 + r\bar{B}_{\mathbb{E}})$ is convex. Then, it suffices to show that y_0 is **absorbing** for $T(x_0 + r\bar{B}_{\mathbb{E}})$; see (1.14). Let $p \in \mathbb{F}$. Since $y_0 \in \text{int } \mathcal{R}(T)$, one can find an $\alpha > 0$ such that $y_0 + \alpha p \in \mathcal{R}(T)$. This implies that there exists $d_\alpha \in \mathbb{E}$ such that

$$y_0 + \alpha p \in T(x_0 + d_\alpha).$$

This d_α is not necessarily in $r\bar{B}_{\mathbb{E}}$. By the convexity of T , one can scale αp and d_α in the same proportion. Indeed, because $y_0 \in T(x_0)$, we also have for all $t \in [0, 1]$ (see exercise ??):

$$(1-t)y_0 + t(y_0 + \alpha p) \in T((1-t)x_0 + t(x_0 + d_\alpha)).$$

Hence, $y_0 + t\alpha p \in T(x_0 + td_\alpha) \subseteq T(x_0 + r\bar{B}_{\mathbb{E}})$, for $t > 0$ small enough.

[(ii) \Rightarrow (iii)] By (ii), there exist radiuses $\alpha > 0$ and $\beta > 0$ such that

$$y_0 + \beta \bar{B}_{\mathbb{F}} \subseteq T(x_0 + \alpha \bar{B}_{\mathbb{E}}). \quad (2.17)$$

The question now is to know whether such an inclusion is maintained for pairs $(x, y) \in \mathcal{G}(T)$ near (x_0, y_0) and for variable radiuses. Let us show that one can take

$$\begin{cases} W = (x_0 + \alpha \bar{B}_{\mathbb{E}}) \times (y_0 + \frac{\beta}{2} \bar{B}_{\mathbb{F}}) \\ \rho = \frac{\beta}{4\alpha} \\ r_{\max} = 2\alpha. \end{cases}$$

Indeed, let $(x, y) \in W \cap \mathcal{G}(T)$. Observe that, by the definition of W and (2.17):

$$y + \frac{\beta}{2} \bar{B}_{\mathbb{F}} \subseteq y_0 + \beta \bar{B}_{\mathbb{F}} \subseteq T(x_0 + \alpha \bar{B}_{\mathbb{E}}). \quad (2.18)$$

Then, for any $t \in [0, 1]$, there holds

$$\begin{aligned} y + t \frac{\beta}{2} \bar{B}_{\mathbb{F}} &= (1-t)y + t(y + \frac{\beta}{2} \bar{B}_{\mathbb{F}}) \\ &\subseteq T((1-t)x + t(x_0 + \alpha \bar{B}_{\mathbb{E}})) \quad [T \text{ convex, } y \in T(x), (2.18)] \\ &= T(x + t((x_0 - x) + \alpha \bar{B}_{\mathbb{E}})) \\ &\subseteq T(x + 2t\alpha \bar{B}_{\mathbb{E}}) \quad [x_0 \in x + \alpha \bar{B}_{\mathbb{E}}]. \end{aligned}$$

Making the change of variable $t \rightsquigarrow r = 2t\alpha$, which is indeed in $[0, r_{\max}]$, we get the desired inclusion since then $t\beta/2 = r\beta/(4\alpha) = \rho r$.

[(iii) \Rightarrow (iv)] One can assume that the neighborhood W of (x_0, y_0) and the maximal radius r_{\max} of the definition 2.13 satisfy

$$W := (x_0, y_0) + \varepsilon(\bar{B}_{\mathbb{E}} \times \bar{B}_{\mathbb{F}}) \quad \text{and} \quad r_{\max} \leq \frac{\varepsilon}{2\rho}, \quad (2.19)$$

for some $\varepsilon > 0$. Let us show that the metric regularity holds at (x_0, y_0) with the module $\mu = 1/\rho$ in the neighborhood W' of (x_0, y_0) defined as follows

$$W' = (x_0, y_0) + \varepsilon'(\bar{B}_{\mathbb{E}} \times \bar{B}_{\mathbb{F}}), \quad \text{with } \varepsilon' \leq \min\left(\varepsilon, \frac{\rho r_{\max}}{1 + \rho}\right). \quad (2.20)$$

Clearly $W' \subseteq W$, since $\varepsilon' \leq \varepsilon$. Let $(x, y) \in W'$. It suffices to show that the inequality (2.16) holds with $\mu = 1/\rho$ for this pair (x, y) . We consider two complementary cases, illustrated in figure 2.3.

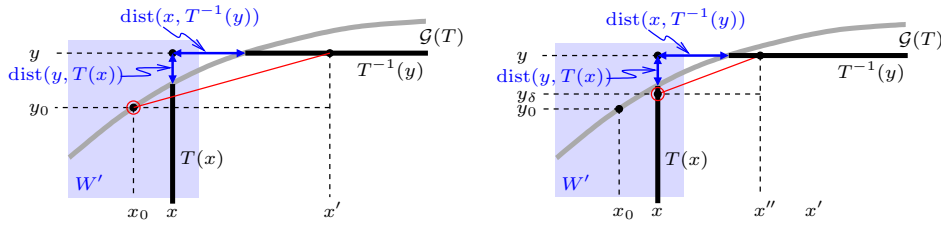


Fig. 2.3. Illustration of the proof of the implication (iii) \Rightarrow (iv) of theorem 2.15: (2.15) is applied at (x_0, y_0) in the left-hand side picture (case 1) and at (x, y_δ) in the right-hand side picture (case 2).

First case, which works when y is far enough from $T(x)$ in the sense that

$$\text{dist}(y, T(x)) \geq (1 + \rho)\varepsilon'. \quad (2.21)$$

One can apply (2.15) at $(x_0, y_0) \in \mathcal{G}(T) \cap W$ (red circle in the left-hand-side picture in figure 2.3), which yields

$$\forall r \in [0, r_{\max}] : y_0 + \rho r \bar{B}_{\mathbb{F}} \subseteq T(x_0 + r \bar{B}_{\mathbb{E}}). \quad (2.22)$$

Note that $y \in y_0 + \rho r \bar{B}_{\mathbb{F}}$, since $\|y - y_0\| \leq \varepsilon' \leq \rho r_{\max}$ ($(x, y) \in W'$ and $\varepsilon' \leq \rho r_{\max}$ by (2.20)). One can also write

$$y = y_0 + \underbrace{\rho \frac{\|y - y_0\|}{\rho}}_{=: r} \underbrace{\begin{cases} \frac{y - y_0}{\|y - y_0\|} & \text{if } y \neq y_0 \\ 0 & \text{otherwise,} \end{cases}}_{\in \bar{B}_{\mathbb{F}}}$$

which shows that $y \in y_0 + \rho r \bar{B}_{\mathbb{F}}$ with $r := \|y - y_0\|/\rho \leq \varepsilon'/\rho \leq r_{\max}$. Therefore, by (2.22), there exists a point $x' \in x_0 + r \bar{B}_{\mathbb{E}}$ such that $y = T(x')$. We can now estimate the distance $\text{dist}(x, T^{-1}(y))$ as follows

$$\begin{aligned} \text{dist}(x, T^{-1}(y)) &\leq \|x - x'\| \quad [x' \in T^{-1}(y)] \\ &\leq \|x - x_0\| + \|x_0 - x'\| \quad [\text{triangular inequality}] \\ &\leq \|x - x_0\| + \frac{1}{\rho} \|y - y_0\| \quad [\|x' - x_0\| \leq r = \|y - y_0\|/\rho] \\ &\leq \varepsilon' + \frac{1}{\rho} \varepsilon' \quad [\text{choice of } x \text{ and } y]. \\ &\leq \frac{1}{\rho} \text{dist}(y, T(x)) \quad [(2.21)], \end{aligned}$$

which is the desired inequality when $\mu = 1/\rho$.

Second case, which works when y is close enough to $T(x)$ in the sense that

$$\text{dist}(y, T(x)) < (1 + \rho)\varepsilon'. \quad (2.23)$$

Then, for any $\delta > 0$ sufficiently small, one can find $y_\delta \in T(x)$ such that

$$\|y - y_\delta\| \leq \text{dist}(y, T(x)) + \delta \quad [\text{definition of the distance}] \quad (2.24)$$

$$\leq (1 + \rho)\varepsilon' \quad [(2.23) \text{ and } \delta > 0 \text{ small}]$$

$$\leq \rho r_{\max} \quad [(2.20)] \quad (2.25)$$

$$\leq \frac{\varepsilon}{2} \quad [(2.19)]. \quad (2.26)$$

One can apply (2.15) at (x, y_δ) (red circle in the right-hand-side picture in figure 2.3), since $(x, y_\delta) \in W \cap G(T)$. Indeed, $(x, y_\delta) \in G(T)$ by construction of $y_\delta \in T(x)$. Next, $\|x - x_0\| \leq \varepsilon' \leq \varepsilon$, by $(x, y) \in W'$ and (2.20). Finally,

$$\|y_\delta - y_0\| \leq \|y_\delta - y\| + \|y - y_0\| \leq \frac{1}{2}\varepsilon + \frac{1}{2}\varepsilon = \varepsilon,$$

where the last inequality comes from (2.26) and from $\|y - y_0\| \leq \varepsilon' \leq \rho r_{\max} \leq \varepsilon/2$ (because $(x, y) \in W'$ and by (2.20) and (2.19)).

The application of (2.15) at (x, y_δ) (red circle in the right-hand-side picture in figure 2.3) yields

$$\forall r \in [0, r_{\max}] : \quad y_\delta + \rho r \bar{B}_{\mathbb{F}} \subseteq T(x + r \bar{B}_{\mathbb{E}}). \quad (2.27)$$

Like in the first case, one can also write

$$y = y_\delta + \underbrace{\rho}_{=:r} \underbrace{\frac{\|y - y_\delta\|}{\|y - y_\delta\|}}_{\in \bar{B}_{\mathbb{F}}} \begin{cases} \frac{y - y_\delta}{\|y - y_\delta\|} & \text{if } y \neq y_\delta \\ 0 & \text{otherwise,} \end{cases}$$

which shows that $y \in y_\delta + \rho r \bar{B}_{\mathbb{F}}$ with $r := \|y - y_\delta\|/\rho \leq r_{\max}$ by (2.25). Therefore, by (2.27), there exists a point $x'' \in x + r \bar{B}_{\mathbb{E}}$ such that $y = T(x'')$. This point x'' allows us to have the following estimation of the distance $\text{dist}(x, T^{-1}(y))$:

$$\begin{aligned} \text{dist}(x, T^{-1}(y)) &\leq \|x - x''\| \quad [x'' \in T^{-1}(y)] \\ &\leq \frac{1}{\rho} \|y - y_\delta\| \quad [x'' \in x + r \bar{B}_{\mathbb{E}} \text{ with } r = \|y - y_\delta\|/\rho] \\ &\leq \frac{1}{\rho} (\text{dist}(y, T(x)) + \delta) \quad [(2.24)]. \end{aligned}$$

Since $\delta > 0$ can be taken arbitrarily small, one deduces the desired inequality when $\mu = 1/\rho$.

[(iv) \Rightarrow (i)] Let us show the contrapositive. If $y_0 \notin \text{int } \mathcal{R}(T)$, one can find a $y \notin \mathcal{R}(T)$ as close as desired to y_0 . For such a y , $T^{-1}(y)$ is empty, implying that $\text{dist}(x_0, T^{-1}(y)) = +\infty$, so that the inequality (2.16) cannot hold for y close to y_0 , since this one would imply that $T(x_0) = \emptyset$, which is in contradiction with the fact that $(x_0, y_0) \in \mathcal{G}(T)$. \square

The goal of property 2.19 below is to show that the **metric regularity** property can “diffuse” from one multifunction to another one close to it, at least for particular forms of multifunctions and perturbations, those in which we are interested in, namely T and T_0 in (2.14). We mean by this that, under certain conditions expressing the proximity of T and T_0 , if T_0 is metric regular, T is also metric regular. It is that property that allows us to have the metric regularity of T from that of T_0 , and reciprocally. The result does not require the convexity of the multifunctions under consideration.

The idea of the proposed proof (there are other possibilities) derives from the following change in perspective [35]. The **metric regularity** gives an upper bound of $\text{dist}(x, T^{-1}(y))$. Observe that

$$x \in T^{-1}(y) \iff y \in T(x) \iff \text{dist}(y, T(x)) = 0,$$

where the last equivalence assumes that $T(x)$ is closed (this assumption is verified if the multifunction T is given by (2.14a) and G is closed). Let us introduce the function $\varphi_y : \mathbb{E} \rightarrow \overline{\mathbb{R}}_+$ defined at $x \in \mathbb{E}$ by

$$\varphi_y(x) := \mu \text{dist}(y, T(x)), \quad (2.28)$$

with $\mu > 0$. We see now that

$$T^{-1}(y) \text{ is the set of the zeros of } \varphi_y. \quad (2.29)$$

The **metric regularity** property can then be expressed as follows:

$$\forall (x, y) \text{ near } (x_0, y_0) : \varphi_y \text{ has a zero in } \overline{B}(x, \varphi_y(x)), \quad (2.30)$$

where $\overline{B}(x, r)$ is the *closed* ball centered at x with radius r .

This viewpoint suggests us to look at conditions ensuring that a function with values in $\overline{\mathbb{R}}_+$ has a zero that is not too far from a given point x . We do this by using two function properties and one lemma.

Definition 2.16 (subcontinuous function) A function $\varphi : \mathbb{E} \rightarrow \overline{\mathbb{R}}_+$ is *subcontinuous* if, for all sequence $\{x_k\} \rightarrow x$ such that $\varphi(x_k) \rightarrow 0$, one has $\varphi(x) = 0$. \square

A continuous function is clearly subcontinuous.

Definition 2.17 (r -steep function) A function $\varphi : \mathbb{E} \rightarrow \overline{\mathbb{R}}_+$ is *r -steep* at x , with $r \in [0, 1)$, if

$$\forall x' \in \overline{B}\left(x, \frac{\varphi(x)}{1-r}\right), \quad \exists x'' \in \overline{B}(x', \varphi(x')) : \varphi(x'') \leq r \varphi(x'). \quad \square$$

One can express this property as follows: for each point x' in the ball $\overline{B}(x, \varphi(x)/(1-r))$, one can find another point x'' , not too far from x' ($x'' \in \overline{B}(x', \varphi(x'))$), at which φ decreases significantly ($\varphi(x'') \leq r \varphi(x')$). Note that, assuming $\varphi(x') \neq 0$,

$$\frac{\varphi(x') - \varphi(x'')}{\|x' - x''\|} \geq \frac{\varphi(x') - r\varphi(x')}{\varphi(x')} = 1 - r,$$

which provides a guarantee on the decrease of φ .

As shown by the next lemma, a function having the two preceding properties has a zero that is not too far from the point at which it is r -steep.

Lemma 2.18 (Lyusternik) *If $\varphi : \mathbb{E} \rightarrow \overline{\mathbb{R}}_+$ is subcontinuous and r -steep at $x \in \text{dom } \varphi$, with $r \in [0, 1)$, then, φ has a zero in $\overline{B}(x, \varphi(x)/(1 - r))$.*

Proof. Let us construct a sequence $\{x_k\}_{k \geq 0} \subseteq \mathbb{E}$ with the following properties:

$$\text{dist}(x_{k+1}, x_k) \leq \varphi(x_k) \quad \text{and} \quad 0 \leq \varphi(x_{k+1}) \leq r\varphi(x_k). \quad (2.31)$$

Take $x_0 = x$. Now, suppose that x_0, \dots, x_k have been determined and let us show how to compute x_{k+1} . Thanks to the properties (2.31), one has

$$\text{dist}(x_k, x) \leq \sum_{i=0}^{k-1} \text{dist}(x_{i+1}, x_i) \leq \sum_{i=0}^{k-1} \varphi(x_i) \leq \left(\sum_{i=0}^{k-1} r^i \right) \varphi(x) \leq \frac{\varphi(x)}{1 - r}. \quad (2.32)$$

Since φ is r -steep at x , one can find $x_{k+1} \in \mathbb{E}$ such that (2.31) holds.

The sequence $\{x_k\}$ is a Cauchy sequence, since $\text{dist}(x_{k+1}, x_k) \leq r^k \varphi(x)$, by (2.31). Therefore, for positive integers $p < q$, there holds

$$\text{dist}(x_q, x_p) \leq \left(\sum_{k=p}^q r^k \right) \varphi(x),$$

where the factor of $\varphi(x)$ in the right-hand side tends to zero when p and $q \rightarrow \infty$. As a result, the sequence $\{x_k\}$ converges: $x_k \rightarrow x \in \overline{B}(x, \varphi(x)/(1 - r))$ by (2.32). Furthermore, taking the limit in $\varphi(x_{k+1}) \leq r\varphi(x_k)$ shows that $\varphi(x_k) \rightarrow 0$. The subcontinuity of φ now implies that $\varphi(x) = 0$. \square

If c is continuous and if the multifunction T is metric regular, then, the function φ_y defined by (2.28) is subcontinuous and 0-steep:

- it is even continuous since $x \mapsto \text{dist}(y, T(x)) = \text{dist}(y, c(x) - G) = \text{dist}(c(x) - y, G)$ is the composition of the continuous functions $x \mapsto c(x) - y$ and $\text{dist}(\cdot, G)$,
- it is 0-steep by the expression (2.30) of the metric regularity.

It looks now reasonable to think that these two properties are maintained for a small perturbation of T . It is the underlying idea of the proof of the next proposition.

Proposition 2.19 (metric regularity diffusion) *let $c : \mathbb{E} \rightarrow \mathbb{F}$ be a continuous function, $\delta : \mathbb{E} \rightarrow \mathbb{F}$ be a Lipschitz continuous function of modulus $L > 0$ in a neighborhood of a point $x_0 \in \mathbb{E}$, and G be a nonempty convex set of \mathbb{F} . If the multifunction*

$$T : \mathbb{E} \rightrightarrows \mathbb{F} : x \mapsto c(x) - G$$

is metric regular at $(x_0, y_0) \in \mathcal{G}(T)$ with modulus $\mu < 1/L$, then the multifunction

$$\tilde{T} : \mathbb{E} \multimap \mathbb{F} : x \mapsto c(x) + \delta(x) - G$$

is metric regular at $(x_0, y_0 + \delta(x_0)) \in \mathcal{G}(\tilde{T})$ with modulus $\mu/(1 - L\mu)$, i.e., for all (x, y) near $(x_0, y_0 + \delta(x_0))$, the following inequality holds

$$\text{dist}(x, \tilde{T}^{-1}(y)) \leq \frac{\mu}{1 - L\mu} \text{dist}(y, \tilde{T}(x)). \quad (2.33)$$

Proof. One can suppose that $\delta(x_0) = 0$ since the metric regularity is invariant with respect to translation of the graph $\mathcal{G}(T)$ in $\mathbb{E} \times \mathbb{F}$. The effects of the continuity are better seen if the function c in T is made visible. One gets, for all $(x, y) \in \mathbb{E} \times \mathbb{F}$:

$$T^{-1}(y) = c^{-1}(G + y), \quad \tilde{T}^{-1}(y) = \tilde{c}^{-1}(G + y), \quad (2.34a)$$

$$\text{dist}(y, T(x)) = \text{dist}(c(x), G + y), \quad \text{dist}(y, \tilde{T}(x)) = \text{dist}(\tilde{c}(x), G + y). \quad (2.34b)$$

The identity (2.34a) comes from the equivalences $x \in T^{-1}(y) \Leftrightarrow y \in T(x) \Leftrightarrow y \in c(x) - G \Leftrightarrow c(x) \in G + y \Leftrightarrow x \in c^{-1}(G + y)$. The identity (2.34b) can be obtained by $\text{dist}(y, T(x)) = \text{dist}(y, c(x) - G) = \text{dist}(c(x), G + y)$.

1) Consider the map

$$\tilde{\varphi}_y : x \in \mathbb{E} \mapsto \tilde{\varphi}_y(x) := \mu \text{dist}(y, \tilde{T}(x)), \quad (2.35)$$

which can be compared to the one introduced in (2.28). Let us show that it suffices to prove the next claim, in which $r := L\mu < 1$:

$$\begin{aligned} \exists U \in \mathcal{N}(x_0), \exists V \in \mathcal{N}(y_0), \forall (x, y) \in U \times V, \\ \tilde{\varphi}_y \text{ is subcontinuous on } \mathbb{E} \text{ and } r\text{-steep at } x. \end{aligned} \quad (2.36)$$

Indeed, by Lyusternik's lemma (lemma 2.18), $\tilde{\varphi}_y$ has then a zero in the ball $\bar{B}(x, \tilde{\varphi}_y(x)/(1 - r))$. Yet, the set of zeros of $\tilde{\varphi}_y$ is $\tilde{T}^{-1}(y)$ (by the same reasoning as the one yielding (2.29); it is here that we use the closure of G). As a result, for all $(x, y) \in U \times V$, one gets (2.33):

$$\text{dist}(x, \tilde{T}^{-1}(y)) \leq \frac{\tilde{\varphi}_y(x)}{1 - r} = \frac{\mu}{1 - L\mu} \text{dist}(y, \tilde{T}(x)).$$

2) It remains to prove (2.36).

By the continuity of c and the continuity of the distance to a convex set, one sees that $\tilde{\varphi}_y$ is continuous on \mathbb{E} , hence subcontinuous on \mathbb{E} .

Let us now determine the neighborhoods U of x_0 and V of y_0 such that, for all $(x, y) \in U \times V$, $\tilde{\varphi}_y$ is r -steep at x .

- By the μ -metric regularity of T at (x_0, y_0) and by (2.34), one can find the neighborhoods U_1 of x_0 and V_1 of y_0 such that, for all $(x, y) \in U_1 \times V_1$, there holds

$$\text{dist}(x, c^{-1}(G + y)) \leq \mu \text{dist}(c(x), G + y). \quad (2.37a)$$

- One can also find the neighborhoods $U_2 \subseteq U_1$ of x_0 and $V_2 \subseteq V_1$ of y_0 such that, for all $(x, y) \in U_2 \times V_2$, there holds

$$\bar{B}\left(x, \mu \operatorname{dist}(\tilde{c}(x), G + y)\right) \subseteq U_1 \quad \text{and} \quad y - \delta(x) \in V_1. \quad (2.37b)$$

The inclusion in U_1 comes from the continuity of $(x, y) \mapsto \operatorname{dist}(\tilde{c}(x), G + y)$ and its vanishment at $(x_0, y_0) \in \mathcal{G}(T)$; the belonging in V_1 comes from the continuity of the map $(x, y) \mapsto y - \delta(x)$ and its value y_0 at (x_0, y_0) .

- Finalement, one takes neighborhoods $U \subseteq U_2$ of x_0 and $V \subseteq V_2$ of y_0 such that, for all $(x, y) \in U \times V$, there holds

$$\bar{B}\left(x, \frac{\mu}{1-r} \operatorname{dist}(\tilde{c}(x), G + y)\right) \subseteq U_2. \quad (2.37c)$$

This is again possible by continuity and vanishment of $\operatorname{dist}(\tilde{c}(x), G + y)$ at $(x, y) = (x_0, y_0)$ (like above).

Let us show that, for all $(x, y) \in U \times V$, $\tilde{\varphi}_y$ is r -steep at x . Let us fix $(x, y) \in U \times V$. Let $x' \in \bar{B}(x, \tilde{\varphi}_y(x)/(1-r))$. One must now find $x'' \in \bar{B}(x', \tilde{\varphi}_y(x'))$ such that $\tilde{\varphi}_y(x'') \leq r\tilde{\varphi}_y(x')$.

- Since $x' \in \bar{B}(x, \tilde{\varphi}_y(x)/(1-r))$, the definition (2.35) of $\tilde{\varphi}_y$ and (2.37c) show that $x' \in U_2$.
- Since $y \in V \subseteq V_2$, one can apply (2.37b) at $(x, y) \rightsquigarrow (x', y)$, which yields

$$\bar{B}\left(x', \mu \operatorname{dist}(\tilde{c}(x'), G + y)\right) \subseteq U_1 \quad \text{and} \quad y' := y - \delta(x') \in V_1. \quad (2.38)$$

- One can therefore use (2.37a) at $(x, y) \rightsquigarrow (x', y')$, which yields

$$\operatorname{dist}(x', c^{-1}(y' + G)) \leq \mu \operatorname{dist}(c(x'), y' + G).$$

Since $c^{-1}(y' + G)$ is closed, this implies that there exists

$$x'' \in c^{-1}(y' + G) \quad (2.39)$$

such that

$$\begin{aligned} \operatorname{dist}(x', x'') &\leq \mu \operatorname{dist}(c(x'), y' + G) \\ &= \mu \operatorname{dist}(c(x'), y - \delta(x') + G) \quad [y' = y - \delta(x') \text{ by (2.38)}] \\ &= \mu \operatorname{dist}(\tilde{c}(x'), G + y) \quad [\tilde{c} = c + \delta] \\ &= \mu \operatorname{dist}(y, \tilde{T}(x')) \quad [(2.34b)] \\ &= \tilde{\varphi}_y(x') \quad [(2.35)]. \end{aligned} \quad (2.40)$$

Therefore, we have found a point x'' at an appropriate distance from x' . We still have to show that this point provides an appropriate decrease of $\tilde{\varphi}_y$.

- Let us use (2.39), which reads

$$c(x'') \in y' + G \quad \text{or} \quad c(x'') + \delta(x'') \in G + y. \quad (2.41)$$

Then,

$$\begin{aligned}
\text{dist}(y, \tilde{T}(x'')) &= \text{dist}(\tilde{c}(x''), G + y) && [(2.34b)] \\
&\leq \text{dist}(c(x'') + \delta(x''), c(x'') + \delta(x')) && [(2.41)] \\
&= \text{dist}(\delta(x''), \delta(x')) \\
&\leq L \text{dist}(x'', x') \\
&\leq L \tilde{\varphi}_y(x') && [(2.40)].
\end{aligned}$$

If the two extreme sides are multiplied by μ , we get thanks to (2.35):

$$\tilde{\varphi}_y(x'') \leq L\mu\tilde{\varphi}_y(x') = r\tilde{\varphi}_y(x').$$

This is the expected inequality. \square

Proof of proposition 2.10. Let T_0 be the multifunction defined by (2.14b). Then, we have seen that

$$\text{(CQ-R) at } x_0 \iff 0 \in \text{int } \mathcal{R}(T_0).$$

The multifunction T is **convex**, since $x \mapsto c(x_0) + c'(x_0) \cdot (x - x_0)$ is affine and G is a convex set, and $(x_0, 0) \in \mathcal{G}(T_0)$, since $x_0 \in X_G$. Then, one can apply theorem 2.15, whose equivalence (i) \Leftrightarrow (iv) yields

$$\text{(CQ-R) at } x_0 \iff T_0 \text{ is } \mu_0\text{-metric regular at } (x_0, 0).$$

We now apply proposition 2.19, with $T \rightsquigarrow T_0$, $\tilde{T} \rightsquigarrow T$ (the multifunction defined in (2.14a)) and $\delta : \mathbb{E} \rightarrow \mathbb{F}$ is defined at $x \in \mathbb{E}$ by

$$\delta(x) = c(x) - c(x_0) - c'(x_0) \cdot (x - x_0).$$

Since c is \mathcal{C}^1 in a neighborhood of x_0 , δ is Lipschitz continuous in a neighborhood of x_0 :

$$\begin{aligned}
\|\delta(x'') - \delta(x')\| &= \|c(x'') - c(x') - c'(x_0) \cdot (x'' - x')\| \\
&\leq \left(\sup_{z \in [x', x'']} \|c'(z) - c'(x_0)\| \right) \|x'' - x'\| \\
&\leq L \|x'' - x'\|,
\end{aligned}$$

where $L > 0$ can be taken as small as desired, provided x' and x'' are sufficiently close to x_0 (c' is continuous), in particular $< 1/\mu_0$. Then, proposition 2.19 shows that the metric regularity of T_0 at $(x_0, 0)$ implies that of T at the same point. One can reverse the roles of T_0 and T , since the convexity of one of the multifunctions does not intervene in proposition 2.19 and L can be taken arbitrarily small by reducing the neighborhood of x_0 . As a result, the metric regularity of T at $(x_0, 0)$ implies that of T_0 at the same point. We have shown that

$$\text{(CQ-R) at } x_0 \iff T \text{ is } \mu\text{-metric regular at } (x_0, 0).$$

It remains to give an interpretation to the right-hand side of the previous equivalence. This one actually means that, for all (x, y) near $(x_0, 0)$, the following holds

$$\text{dist}(x, T^{-1}(y)) \leq \mu \text{dist}(y, T(x)).$$

Since $\text{dist}(x, T^{-1}(y)) = \text{dist}(x, c^{-1}(G + y))$ and $\text{dist}(y, T(x)) = \text{dist}(c(x), G + y)$, the equivalence of proposition 2.10 is proven. \square

2.1.4 Robinson's Constraint Qualification

Let us recall the *Robinson condition* (2.42) and write it at a point $x \in \mathbb{E}$:

$$(CQ-R) \quad 0 \in \text{int}(c(x) + c'(x)\mathbb{E} - G). \quad (2.42)$$

The main goal of this section is to show that this condition is sufficient for ensuring that the constraint qualification conditions (2.6) hold at a given point of X_G . To achieve this goal, it is useful to have at hand the few equivalent formulations of (CQ-R), given in the next proposition. Recall some definitions: for a closed convex set C and a point $x \in C$, the **cone of feasible directions** to C at x reads $\mathbf{T}_x^f C := \mathbb{R}_+(C - x)$, while the **tangent cone** $\mathbf{T}_x C$ to C at x is the closure of $\mathbf{T}_x^f C$.

Proposition 2.20 (other formulations of (CQ-R)) *If c is differentiable at $x \in X_G$, the following properties are equivalent:*

$$0 \in \text{int}(c(x) + c'(x)\mathbb{E} - G), \quad (2.43a)$$

$$c'(x)\mathbb{E} - \mathbf{T}_{c(x)}^f G = \mathbb{F}, \quad (2.43b)$$

$$c'(x)\mathbb{E} - \mathbf{T}_{c(x)} G = \mathbb{F}, \quad (2.43c)$$

$$\overline{c'(x)\mathbb{E} - \mathbf{T}_{c(x)} G} = \mathbb{F}. \quad (2.43d)$$

Proof. [(2.43a) \Rightarrow (2.43b)] It suffices to prove the inclusion “ \supseteq ”. Let $y \in \mathbb{F}$. Then, by (2.43a), $y/t \in c'(x)\mathbb{E} - (G - c(x))$ for a sufficiently large $t > 0$. Hence $y \in c'(x)[t\mathbb{E}] - [t(G - c(x))] \subseteq c'(x)\mathbb{E} - \mathbf{T}_{c(x)}^f G$.

[(2.43b) \Rightarrow (2.43c) \Rightarrow (2.43d)] These implications are clearly satisfied, since the sets in the left-hand sides are larger and larger.

[(2.43c) \Rightarrow (2.43b)] We start by exploiting the link between $\mathbf{T}_{c(x)}^f G$ and $\mathbf{T}_{c(x)} G$:

$$\begin{aligned} c'(x)\mathbb{E} - \mathbf{T}_{c(x)} G &= c'(x)\mathbb{E} - \overline{\mathbf{T}_{c(x)}^f G} && \text{[definition of } \mathbf{T}_{c(x)} G \text{]} \\ &\subseteq \overline{c'(x)\mathbb{E} - \mathbf{T}_{c(x)}^f G} && \text{[} \overline{A + B} \subseteq \overline{A} + \overline{B} \text{]} \\ &\subseteq \overline{c'(x)\mathbb{E} - \mathbf{T}_{c(x)} G} && \text{[} \mathbf{T}_{c(x)}^f G \subseteq \mathbf{T}_{c(x)} G \text{]} \\ &= \mathbb{F} && \text{[(2.43d)].} \end{aligned}$$

Taking the closure of all these sets shows that

$$\overline{c'(x)\mathbb{E} - \mathbf{T}_{c(x)}^f G} = \mathbb{F}.$$

Since $c'(x)\mathbb{E} - \mathbf{T}_{c(x)}^f G$ is convex, this identity certainly implies (2.43b) (see also exercise 1.2.4).

[(2.43b) \Rightarrow (2.43a)] From (1.14), it suffices to prove that 0 is **absorbing** for the convex set $C := c(x) + c'(x)\mathbb{E} - G$ de \mathbb{F} . Let $y \in \mathbb{F}$. Condition (2.43b) implies that there exists $d \in \mathbb{E}$, $y_1 \in G$, and $\alpha > 0$ such that $y = c'(x)d - \alpha(y_1 - c(x))$. For $t := 1/\alpha$, we get $ty = c(x) + c'(x)(td) - y_1 \in C$; hence 0 is absorbing for C . \square

Note that, in (2.43b)-(2.43d), one can change the minus sign by a plus, since $c'(x)\mathbb{E}$ and \mathbb{F} are vector spaces. Conditions (2.43b) and (2.43c) convey the fact that \mathbb{F} can be written as the sum of a subspace, say \mathbb{F}_0 , and a cone, say K . According to exercise 1.2.7, this is equivalent to saying that

$$\mathbb{F}_0 + \text{vect } K = \mathbb{F} \quad \text{and} \quad \mathbb{F}_0 \cap (\text{ri } K) \neq \emptyset.$$

Proposition 2.21 (constraint qualification by (CQ-R)) *If c is continuously differentiable in the neighborhood of a point $x \in X_G$ and if the Robinson condition (CQ-R) holds at this point, then, the constraint c is qualified to represent X_G at x , in the sense of definition 2.4.*

Proof. We must show that (CQ-R) implies the two properties (2.6a) and (2.6b).

[(2.6a)] For this step of the proof, it suffices to use Robinson's error bound (2.12). According to proposition 2.3, it suffices to show the inclusion $T'_x X_G \subseteq T_x X_G$. Let $d \in T'_x X_G$. To show that $d \in T_x X_G$, it suffices to build a sequence $\{x_k\} \subseteq X_G$ (a priori a difficult task) and a sequence $\{t_k\} \downarrow 0$ such that $(x_k - x)/t_k \rightarrow d$. The followed approach is illustrated by figure 2.4: thanks to the Robinson error bound (2.12), one

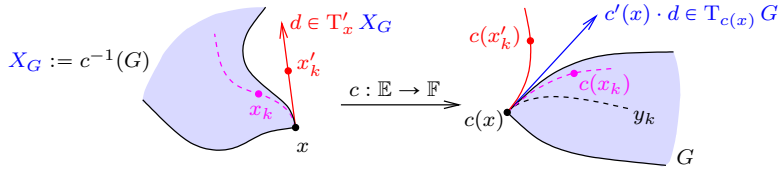


Fig. 2.4. Illustration of the proof of (2.6a) to establish proposition 2.21

can create a sequence $\{x_k\} \subseteq X_G := c^{-1}(G)$ with the desired properties; these are due to the good feature of the sequence $\{c(x + t_k d)\}$ in \mathbb{F} , which is to be asymptotically very close to the sequence $\{y_k\}$. Here are the details.

Since d is in the **linearizing cone** $T'_x X_G$, it satisfies $c'(x) \cdot d \in T_{c(x)} G$, which implies that there exist sequences $\{y_k\} \subseteq G$ and $\{t_k\} \downarrow 0$ (it is the good one!), such that

$$\frac{y_k - c(x)}{t_k} \rightarrow c'(x) \cdot d \quad \text{or} \quad c(x) + c'(x) \cdot (t_k d) - y_k = o(t_k). \quad (2.44)$$

Consider the sequence $\{x'_k\} \subseteq \mathbb{E}$ defined by

$$x'_k := x + t_k d.$$

This sequence is not necessary in X_G ; hence the sequence $\{c(x'_k)\}$ is not necessary in G , but it is not far from $\{y_k\}$ (which is in G) asymptotically. Indeed, by the differentiability of c , one has

$$c(x'_k) = c(x) + c'(x) \cdot (t_k d) + o(t_k)$$

and therefore, using (2.44),

$$c(x'_k) = y_k + o(t_k). \quad (2.45)$$

We now build the appropriate sequence $\{x_k\} \subseteq X_G$ thanks to Robinson's error bound (2.12) at $x_0 \rightsquigarrow x$. This one can be applied at x'_k since this point is close to x for sufficiently large k (and c is C^1 in a neighborhood of x). It tells us that there exists a constant $\mu \geq 0$ such that, for sufficiently large k , one has

$$\begin{aligned} \text{dist}(x'_k, X_G) &\leq \mu \text{dist}(c(x'_k), G) && [(2.12)] \\ &\leq \mu \text{dist}(c(x'_k), y_k) && [y_k \in G] \\ &= o(t_k) && [(2.45)]. \end{aligned}$$

This implies that there exists a sequence $\{x_k\} \subseteq X_G$ such that $x_k = x'_k + o(t_k)$. We have build the desired sequence $\{x_k\}$ since

$$\frac{x_k - x}{t_k} = \frac{x_k - x'_k}{t_k} + \frac{x'_k - x}{t_k} = \frac{o(t_k)}{t_k} + d \rightarrow d.$$

[(2.6b)] This step of the proof uses the expression (2.43c) of (CQ-R). We consider a convergent sequence in the cone $\mathcal{K} := c'(x)^*[(T_{c(x)}G)^+]$, whose closure has to be shown:

$$c'(x)^*\lambda_k \rightarrow y, \quad (2.46)$$

where $\{\lambda_k\} \subseteq (T_{c(x)}G)^+$. The goal now is to show that y is in \mathcal{K} . In finite dimension, it suffices to show that the sequence $\{\lambda_k\}$ is bounded (then, one extracts from it a convergent subsequence, whose limit is in the closed cone $(T_{c(x)}G)^+$, and one takes the limit in (2.46), which shows that $y \in \mathcal{K}$). One proceeds by contradiction. If $\{\lambda_k\}$ is unbounded, one can extract a subsequence from $\{\lambda_k/\|\lambda_k\|\} \subseteq (T_{c(x)}G)^+$, which converges to a *nonzero* vector (it has unit norm), say μ . We see on (2.46) that μ verifies

$$c'(x)^*\mu = 0 \quad \text{and} \quad \mu \in (T_{c(x)}G)^+. \quad (2.47)$$

Since $\mu \in \mathbb{F}$, (2.43c) tells us that it can be written $\mu = c'(x)d - p$ for some $d \in \mathbb{E}$ and $p \in T_{c(x)}G$. Therefore,

$$\|\mu\|^2 = \langle \mu, c'(x)d - p \rangle = -\langle \mu, p \rangle \leq 0,$$

where we have used (2.47) and $p \in T_{c(x)}G$. This implies that $\mu = 0$, which is in contradiction with the fact observed above, indicating that $\mu \neq 0$. \square

2.1.5 Set of optimal multipliers

In the light of the optimality conditions (2.8a)-(2.8b), the set of optimal multipliers associated with a stationary point x of problem (P_G) reads

$$\Lambda = \{\lambda \in N_{c(x)}G : \nabla f(x) + c'(x)^*\lambda = 0\}, \quad (2.48)$$

where we have alleviated notation by dropping the star indices. As an intersection of a closed convex cone and an affine subspace, this is a closed convex set. This section gives more properties of this set.

The next proposition cares about the boundedness of Λ (hence of its compactness) and, in this sense, extends point 1 of proposition 1.42 from problem (P_{EI}) to problem (P_G) . It does it by computing first its asymptotic cone Λ^∞ and then uses proposition 1.13 to characterize the boundedness of Λ .

Proposition 2.22 (boundedness of the set of multipliers, Gauvin's property) *Suppose that f and c are differentiable at a stationary point x of problem (P_G) and that the set Λ of associated multipliers is nonempty. Then, the asymptotic cone of Λ reads*

$$\Lambda^\infty = (\mathcal{N}_{c(x)} G) \cap \mathcal{N}(c'(x)^*) = [c'(x)\mathbb{E} - \mathcal{T}_{c(x)} G]^+. \quad (2.49)$$

It results that, Λ is bounded if and only if the Robinson condition (CQ-R) holds at $x_0 = x$.

Proof. According to exercise 1.2.13, the asymptotic cone of Λ is given by the first identity in (2.49). Now, using point 4 of proposition 1.7, $\mathcal{N}_{c(x)} G = (-\mathcal{T}_{c(x)} G)^+$, and $\mathcal{N}(c'(x)^*) = \mathcal{R}(c'(x))^\perp = (c'(x)\mathbb{E})^+$, we get the second identity in (2.49).

Next, according to proposition 1.13, Λ is bounded if and only if its asymptotic cone is reduced to $\{0\}$. Hence, we only have to prove

$$[c'(x)\mathbb{E} - \mathcal{T}_{c(x)} G]^+ = \{0\} \quad \iff \quad \text{(CQ-R)}. \quad (2.50)$$

The implication “ \Leftarrow ” is clear, if we consider the form (2.43c) of (CQ-R). For the reverse implication “ \Rightarrow ”, take the dual of the two sides of the identity in the left-hand side of (2.50) and use (1.16) to get $\text{cl}(c'(x)\mathbb{E} - \mathcal{T}_{c(x)} G) = \mathbb{F}$, which is equivalent to (CQ-R) by (2.43d). \square

The final claim of the previous proposition may look strange since (CQ-R) only depends on the constraints of the optimization problem (P_G) , while the optimal multiplier set Λ defined by (2.48), which is used to characterize (CQ-R), also depends on the objective of the problem. Observe, however, that the characterization only uses the asymptotic cone Λ^∞ , which does not depend on the objective of the optimization problem.

Proposition 2.24 [129] below highlights conditions ensuring that Λ is a singleton, which is an esoteric way of saying that there is a unique multiplier associated with the given stationary point x . We pave the way to this uniqueness result by the next lemma, which gives an expression of the dual cone of $c'(x)\mathbb{E} - [(\mathcal{T}_{c(x)} G) \cap \lambda^\perp]$, where, for $\lambda \in \mathbb{F}$:

$$\lambda^\perp := \{\mu \in \mathbb{F} : \langle \mu, \lambda \rangle = 0\} = (\mathbb{R}\{\lambda\})^\perp.$$

Note that the a priori larger set $c'(x)\mathbb{E} - \mathcal{T}_{c(x)} G$ is equal to \mathbb{F} , when (CQ-MF) holds; see proposition 2.20. A comparison with the dual cone of $c'(x)\mathbb{E} - \mathcal{T}_{c(x)} G$, whose expression is given in (2.49), may be instructive.

Lemma 2.23 *Let $x \in \mathbb{E}$ be a feasible point of problem (P_G) and let $\lambda \in N_{c(x)} G$. Then,*

$$[(T_{c(x)} G) \cap \lambda^\perp]^+ = -T_\lambda(N_{c(x)} G), \quad (2.51a)$$

$$(c'(x)\mathbb{E} - [(T_{c(x)} G) \cap \lambda^\perp]^+)^+ = \mathcal{N}(c'(x)^*) \cap T_\lambda(N_{c(x)} G). \quad (2.51b)$$

Proof. [(2.51a)] Let us first compute

$$\begin{aligned} [(T_{c(x)} G) \cap \lambda^\perp]^+ &= \overline{(T_{c(x)} G)^+ + \{\lambda^\perp\}^+} \quad [(1.21)] \\ &= \overline{-N_{c(x)} G + \mathbb{R}\{\lambda\}} \quad [(T_{c(x)} G)^+ = -N_{c(x)} G] \\ &= -T_\lambda^f(N_{c(x)} G) \quad [\lambda \in N_{c(x)} G \text{ and exercise 1.2.11}] \\ &= -T_\lambda(N_{c(x)} G) \quad [(1.22)]. \end{aligned}$$

[(2.51b)] Using property 6 of proposition 1.7, one gets

$$(c'(x)\mathbb{E} - [(T_{c(x)} G) \cap \lambda^\perp]^+)^+ = (c'(x)\mathbb{E})^+ \cap -[(T_{c(x)} G) \cap \lambda^\perp]^+.$$

One can then conclude by using $(c'(x)\mathbb{E})^+ = \mathcal{R}(c'(x))^\perp = \mathcal{N}(c'(x)^*)$, since $c'(x)\mathbb{E} = \mathcal{R}(c'(x))$ is a subspace, and (2.51a). \square

Proposition 2.24 (uniqueness of the optimal multiplier) *Let $(x, \lambda) \in \mathbb{E} \times \mathbb{F}$ be a stationary pair of problem (P_G) and denote by Λ the set (2.48) of optimal multipliers associated with x . Consider the following properties:*

- (i) $\Lambda = \{\lambda\}$,
- (ii) $\mathcal{N}(c'(x)^*) \cap T_\lambda^f(N_{c(x)} G) = \{0\}$,
- (iii) $\mathcal{N}(c'(x)^*) \cap T_\lambda(N_{c(x)} G) = \{0\}$,
- (iv) $c'(x)\mathbb{E} - [(T_{c(x)} G) \cap \lambda^\perp] = \mathbb{F}$.

Then, (i) \Leftrightarrow (ii) \Leftrightarrow (iii) \Leftrightarrow (iv). If $T_\lambda^f(N_{c(x)} G) = T_\lambda(N_{c(x)} G)$, then the four properties (i)-(iv) are equivalent.

Proof. [(i) \Rightarrow (ii)] Let $\mu \in \mathcal{N}(c'(x)^*) \cap T_\lambda^f(N_{c(x)} G)$. It suffices to show that $\mu = 0$. Since $\mu \in T_\lambda^f(N_{c(x)} G)$, it is of the form $\mu = \alpha(\lambda' - \lambda)$ for some $\alpha \geq 0$ and $\lambda' \in N_{c(x)} G$. If $\alpha = 0$, we get $\mu = 0$ as desired. Otherwise, $\lambda' = \lambda + \mu/\alpha$, which satisfies

$$\nabla f(x) + c'(x)^* \lambda' = 0 \quad \text{and} \quad \lambda' \in N_{c(x)} G.$$

Hence $\lambda' \in \Lambda = \{\lambda\}$, implying that $\lambda' = \lambda$. Hence $\mu = 0$.

[(i) \Leftarrow (ii)] Let $\lambda' \in \Lambda$ and set $\mu := \lambda' - \lambda$. It suffices to show that $\mu = 0$. It immediately follows from λ and $\lambda' \in \Lambda$ in (2.48) that

$$c'(x)^* \mu = 0 \quad \text{and} \quad \mu \in T_\lambda^f(N_{c(x)} G).$$

Then, (ii) implies that $\mu = 0$.

[(ii) \Leftarrow (iii)] This is because $\mathbf{T}_\lambda^f(\mathbf{N}_{c(x)} G) \subseteq \mathbf{T}_\lambda(\mathbf{N}_{c(x)} G)$.

[(iii) \Rightarrow (iv)] Taking the dual of both sides of the identity in (iii) and using (2.51b), we have that the closure of $c'(x)\mathbb{E} - [(\mathbf{T}_{c(x)} G) \cap \lambda^\perp]$ is \mathbb{F} (see (1.16)). Since this last set is convex, we certainly have (iv) (see exercise 1.2.4).

[(iii) \Leftarrow (iv)] Taking the dual of both sides of the identity in (iv) and using (2.51b), we get (iii).

[Last claim] If $\mathbf{T}_\lambda^f(\mathbf{N}_{c(x)} G) = \mathbf{T}_\lambda(\mathbf{N}_{c(x)} G)$, then (ii) and (iii) are clearly identical and, therefore, all the properties (i)-(iv) are equivalent. \square

Remarks 2.25 1) When G is a convex polyhedron, like in problem (P_{EI}) , its normal cone $\mathbf{N}_{c(x)} G$ is also a convex polyhedron (see (1.27a)). Then, the tangent cone and the cone of feasible directions to $\mathbf{N}_{c(x)} G$ are identical (see (1.26a)) and the four properties (i)-(iv) of proposition 2.24 are equivalent.

2) One can recover the conditions 2.(ii) and 2.(iii) of proposition 1.42 for problem (P_{EI}) from the conditions (iii) and (iv) of the previous proposition. This is part of the subject of exercise 2.1.6.

2.1.6 Other problems \blacktriangle

In this section, we show how the results obtained for the general problem (P_G) can be applied to other problems.

Problem with an additional set constraint \blacktriangle

Let \mathbb{E} and \mathbb{F} be two Euclidean vector spaces. Consider the problem

$$(P_{Q,G}) \quad \begin{cases} \min f(x) \\ x \in Q \\ c(x) \in G, \end{cases} \quad (2.52)$$

where $f : \mathbb{E} \rightarrow \mathbb{R}$ and $c : \mathbb{E} \rightarrow \mathbb{F}$ are two smooth functions, Q is a nonempty closed convex set of \mathbb{E} and G is a nonempty closed convex set of \mathbb{F} . We denote its feasible set by

$$X_{Q,G} := \{x \in \mathbb{E} : x \in Q, c(x) \in G\}.$$

We now give a series of results, whose proofs are proposed in exercise 2.1.8.

We first look at various adaptations to problem $(P_{Q,G})$ of the Robinson constraint qualification assumption (CQ-R) associated with problem (P_G) . They are similar to those given in proposition 2.20.

Proposition 2.26 (constraint qualification) *If c is differentiable at a feasible point $x \in X_{Q,G}$, the following properties are equivalent:*

$$0 \in \text{int}\left(c(x_*) + c'(x_*)(Q - x_*) - G\right), \tag{2.53a}$$

$$\tag{2.53b}$$

$$\tag{2.53c}$$

$$\tag{2.53d}$$

The following result gives necessary first order optimality conditions (NC1) for problem $(P_{Q,G})$.

Proposition 2.27 (NC1 for problem $(P_{Q,G})$) *Let x_* be a local solution to problem $(P_{Q,G})$. Suppose that f and c are differentiable at x_* and that the constraint qualification assumption (2.53) holds. Then, there exists $\lambda_* \in \mathbb{F}$ such that*

$$\nabla f(x_*) + c'(x_*)^* \lambda_* \in (\mathbb{T}_{x_*} Q)^+ \tag{2.54a}$$

$$\lambda_* \in N_{c(x_*)} G. \tag{2.54b}$$

As usual, the NC1 become sufficient optimality conditions of the first order (SC1) for a “convex problem” $(P_{Q,G})$, in a sense specified in the next proposition.

Proposition 2.28 (SC1 for a convex $(P_{Q,G})$) *Suppose that problem $(P_{Q,G})$ is convex in the sense that f is convex and its feasible set $X_{Q,G}$ is convex. Suppose also that f and c are differentiable at a point $x_* \in X_{Q,G}$, and that there exists a multiplier $\lambda_* \in \mathbb{F}$ such that (x_*, λ_*) verifies the first order optimality conditions (2.54). Then, x_* is a global solution to $(P_{Q,G})$.*

One can also adapt, from propositions 2.22 and 2.24, the conditions ensuring the boundedness of the set of optimal multipliers and the uniqueness of the optimal multiplier.

Proposition 2.29 (boundedness of the set of multipliers, Gauvin’s property) *Suppose that f and c are differentiable at a stationary point x of problem $(P_{Q,G})$ and that the set Λ of associated multipliers is nonempty. Then, the asymptotic cone of Λ reads*

$$\Lambda^\infty \tag{2.55}$$

It results that, Λ is bounded if and only if

Proposition 2.30 (uniqueness of the optimal multiplier) *Let $(x, \lambda) \in \mathbb{E} \times \mathbb{F}$ be a stationary pair of problem $(P_{Q,G})$ and denote by Λ the set (2.48) of optimal multipliers associated with x . Consider the following properties:*

(i) $\Lambda = \{\lambda\}$,

(ii)

(iii)

(iv)

Then, (i) \Leftrightarrow (ii) \Leftarrow (iii) \Leftrightarrow (iv).

Composite optimization

Let \mathbb{E} and \mathbb{F} be two Euclidean vector spaces. As already quoted in example 2.2(3), a *composite optimization problem* is an optimization problem of the form

$$(P_{\text{comp}}) \quad \begin{cases} \min_{x \in \mathbb{E}} (g \circ F)(x) \\ x \in Q, \end{cases} \quad (2.56)$$

where $F : \mathbb{E} \rightarrow \mathbb{F}$ is a differentiable function, $g \in \overline{\text{Conv}}(\mathbb{F})$, $(g \circ F)$ denotes the composition of g and F [hence $(g \circ F)(x) = g(F(x))$] and Q is a nonempty closed convex set of \mathbb{E} . These assumptions are made in all this section.

Definition 2.31 (convex problem (P_{comp})) The problem (P_{comp}) is said to be *convex* if the function $(g \circ F) + \mathcal{I}_Q$ is convex or, equivalently, if $g \circ F$ is convex on the convex set Q . \square

Problem (2.56) is equivalent to (same optimal value, identical x -solution)

$$\begin{cases} \min_{(x,\alpha) \in \mathbb{E} \times \mathbb{R}} \alpha \\ x \in Q \\ (g \circ F)(x) \leq \alpha. \end{cases}$$

Since the first constraint also reads $(x, \alpha) \in Q \times \mathbb{R}$ and the second constraint also reads $g(F(x)) \leq \alpha$ or $(F(x), \alpha) \in \text{epi}(g)$, this last problem, and therefore (2.56), is equivalent to solving

$$\begin{cases} \min_{(x,\alpha) \in \mathbb{E} \times \mathbb{R}} \alpha \\ (x, \alpha) \in Q \times \mathbb{R} \\ (F(x), \alpha) \in \text{epi}(g). \end{cases} \quad (2.57)$$

Problem (2.57) is of the form $(P_{Q,G})$ in (2.52), so that the results obtained for this latter problem can be used to determine analog properties for problem (P_{comp}) .

The proof of the propositions of this section are proposed in exercise 2.1.9.

We start by adapting to problem (P_{comp}) the constraint qualification assumptions (2.53) associated with problem (2.57). The reason of the need of such a function qualification might not be obvious when one looks at the formulation (2.56) of the composite problem. Actually, it is linked to implicit constraints introduced by the infinite values that g can take. Observe indeed that the equivalent conditions (2.58) of the next proposition, in particular its condition (2.58c), are satisfied if the considered

point $x \in \mathbb{E}$ is such that $F(x) \in \text{int}(\text{dom } g)$ (since then $\mathbb{T}_{F(x)}(\text{dom } g) = \mathbb{F}$), that is, when the infinite values of g do not appear in the neighborhood of $F(x)$.

Proposition 2.32 (function qualification) *The following properties are equivalent:*

$$0_{\mathbb{F}} \in \text{int} \left(F(x) + F'(x)(Q - x) - (\text{dom } g) \right), \quad (2.58a)$$

$$F'(x)(\mathbb{T}_x^f Q) - \mathbb{T}_{F(x)}^f(\text{dom } g) = \mathbb{F}, \quad (2.58b)$$

$$F'(x)(\mathbb{T}_x Q) - \mathbb{T}_{F(x)}(\text{dom } g) = \mathbb{F}, \quad (2.58c)$$

$$\overline{F'(x)(\mathbb{T}_x Q) - \mathbb{T}_{F(x)}(\text{dom } g)} = \mathbb{F} \quad (2.58d)$$

and, for any $\alpha \geq g(F(x))$, these conditions are equivalent to the *Robinson constraint qualification (CQ-R)* at (x, α) for the constraints of problem (2.57).

Proposition 2.27 for problem (2.57) yields the following necessary optimality conditions of the first order (NC1) for problem (P_{comp}) . Observe that if g is differentiable at $F(x)$, (2.59) becomes

$$\nabla(g \circ F)(x_*) \in (\mathbb{T}_{x_*} Q)^+,$$

which is then the expected Peano-Kantorovich optimality condition (1.51) for problem (P_{comp}) . This observation should help to understand and remember (2.59).

Proposition 2.33 (NC1 for problem (P_{comp})) *Suppose that x_* is a local solution to problem (P_{comp}) and that the function qualification condition (2.58) holds. Then, there exists $\lambda_* \in \mathbb{F}$ such that*

$$F'(x_*)^* \lambda_* \in (\mathbb{T}_{x_*} Q)^+, \quad (2.59a)$$

$$\lambda_* \in \partial g(F(x_*)), \quad (2.59b)$$

where $\partial g(y)$ denotes the *subdifferential* of g at y .

As usual, the NC1 become sufficient conditions of the first order (SC1) for *global* optimality of a convex problem (P_{comp}) . It turns out that the appropriate notion of convexity for problem (P_{comp}) is the one of definition 2.31 (actually, it has been introduced for being useful in the next proposition).

Proposition 2.34 (SC1 for a convex (P_{comp})) *Suppose that problem (P_{comp}) is convex in the sense of definition 2.31, that $x_* \in Q$ and that there exists a multiplier $\lambda_* \in \mathbb{F}$ such that (x_*, λ_*) verifies the first order optimality conditions (2.59). Then, x_* is a global solution to (P_{comp}) .*

Notes

The condition (2.43c) is reminiscent of the notion of transversality in differential geometry [109]. The link between the boundedness of the set of optimal multipliers and (CQ-MF) in nonlinear optimization (point 1 of proposition 1.42, extended to problem (P_G) by proposition 2.22) was observed by Gauvin [59; 1977]. More on metric regularity can be found in [44; 2009].

Exercises

2.1.1. *Convex (P_{EI}) .* Let $G = \{0_{\mathbb{R}^m}^E\} \times \mathbb{R}_-^m$ and $T : \mathbb{R}^n \multimap \mathbb{R}^m : x \mapsto c(x) - G$ be the multifunction associated with the constraint of problem (P_{EI}) . Show that T is convex if and only if c_E is affine and c_I is componentwise convex.

2.1.2. *Open mapping theorem.* Let $A : \mathbb{E} \rightarrow \mathbb{F}$ be a linear map between the finite dimension normed vector spaces \mathbb{E} and \mathbb{F} (the result also holds for infinite dimension Banach spaces, but is more difficult to prove [24]). Denote by $\bar{B}_{\mathbb{E}}$ and $\bar{B}_{\mathbb{F}}$ the closed unit balls of \mathbb{E} and \mathbb{F} respectively. Then, the following properties are equivalent:

- (i) A is surjective,
- (ii) $\exists \rho > 0$ such that $\rho \bar{B}_{\mathbb{F}} \subseteq A(\bar{B}_{\mathbb{E}})$,
- (iii) $\exists \mu > 0$ such that: $\forall y \in \mathbb{F}, \exists x \in \mathbb{E}$ such that $Ax = y$ and $\|x\| \leq \mu \|y\|$.

2.1.3. *Examples of use of Robinson's constraint qualification condition.* Consider the following sets of the form $X_G := \{x \in \mathbb{E} : c(x) \in G\}$, with $G \subseteq \mathbb{F}$, and points $x_0 \in X_G$:

- 1) $\mathbb{E} = \mathbb{F} = \mathbb{R}^2$, and c and G are defined by

$$c(x) = (x_1^2 + (x_2 - 1)^2 - 1, x_1^2 + (x_2 + 1)^2 - 1), \quad G = \mathbb{R}_-^2, \quad \text{and} \quad x_0 = (0, 0),$$

- 2) $\mathbb{E} = \mathbb{F} = \mathbb{R}^2$, and c (2 identical constraints) and G are defined by

$$c(x) = (x_2, x_2), \quad G = \mathbb{R}_+^2, \quad \text{and} \quad x_0 = (0, 0).$$

For these sets X_G and points $x_0 \in X_G$,

- (i) determine whether Robinson's constraint qualification condition holds at x_0 , without using its equivalence with (CQ-MF),
- (ii) find a modulus of metric regularity of the multifunction $x \mapsto c(x) - G$ at $(x_0, 0)$ if any.

2.1.4. *Robinson's constraint qualification condition for (P_{EI}) .* Viewing problem (P_{EI}) as a particular instance of problem (P_G) , show that the Robinson constraint qualification condition (CQ-R) is equivalent to the Mangasarian-Fromovitz's constraint qualification condition (CQ-MF).

2.1.5. *Robinson's constraint qualification and the primal SDO problem.* Let \mathcal{S}^n be the set of symmetric matrices of order n , which is equipped with the scalar product $\langle A, B \rangle = \text{tr } AB$ (the trace of AB). Let \mathcal{S}_+^n be the cone of \mathcal{S}^n made of the positive semidefinite matrices and \mathcal{S}_{++}^n be the cone of \mathcal{S}^n made of the positive definite matrices. We abbreviate $X \succeq 0$ for $X \in \mathcal{S}_+^n$ and $X \succ 0$ for $X \in \mathcal{S}_{++}^n$.

Consider the primal semidefinite optimization (SDO) problem, written as follows

$$(P) \quad \begin{cases} \inf_{X \in \mathcal{S}^n} \langle C, X \rangle \\ \mathcal{A}(X) = b \\ X \succeq 0, \end{cases} \quad (2.60)$$

where $C \in \mathcal{S}^n$, $\mathcal{A} : \mathcal{S}^n \rightarrow \mathbb{R}^m$ is a linear map, and $b \in \mathbb{R}^m$. The feasible and strictly feasible sets of this problem are respectively denoted by

$$\mathcal{F}_P := \{X \in \mathcal{S}_+^n : \mathcal{A}(X) = b\} \quad \text{and} \quad \mathcal{F}_P^s := \{X \in \mathcal{S}_{++}^n : \mathcal{A}(X) = b\}.$$

We assume that $\mathcal{F}_P \neq \emptyset$.

Let us represent the feasible set of problem (2.60) by $\mathcal{F}_P := \{X \in \mathcal{S}^n : c(X) \in G\}$, where

$$c : X \in \mathcal{S}^n \mapsto (\mathcal{A}(X) - b, X) \in \mathbb{R}^m \times \mathcal{S}^n \quad \text{and} \quad G = \{0_{\mathbb{R}^m}\} \times \mathcal{S}_+^n. \quad (2.61)$$

Let (CQ-R) denote Robinson's constraint qualification condition for the above representation of \mathcal{F}_P .

1) Show that (CQ-R) at $X_0 \in \mathcal{F}_P$ can equivalently be written

$$(\mathcal{A}, I)\mathcal{S}^n - \{0_{\mathbb{R}^m}\} \times T_{X_0} \mathcal{S}_+^n = \mathbb{R}^m \times \mathcal{S}^n, \quad (2.62)$$

where $(\mathcal{A}, I) : \mathcal{S}^n \rightarrow \mathbb{R}^m \times \mathcal{S}^n$ is the linear map $X \mapsto (\mathcal{A}(X), X)$.

2) Show that if (CQ-R) holds at *some* point $X_0 \in \mathcal{F}_P$, then

$$\mathcal{A} \text{ is surjective} \quad \text{and} \quad \mathcal{F}_P^s \neq \emptyset. \quad (2.63)$$

3) Show that if (2.63) holds then (CQ-R) holds at *any* point $X_0 \in \mathcal{F}_P$.

4) Show that if X solves (P) and (2.63) holds then there is a pair $(y, S) \in \mathbb{R}^m \times \mathcal{S}^n$ such that

$$\mathcal{A}^*(y) + S = C, \quad S \succeq 0, \quad \text{and} \quad \langle X, S \rangle = 0. \quad (2.64)$$

Furthermore, the set of pairs $(y, S) \in \mathbb{R}^m \times \mathcal{S}^n$ satisfying these three conditions is compact.

5) Reciprocally, show that if $X \in \mathcal{F}_P$ is such that the set of pairs (y, S) verifying (2.64) is nonempty and bounded, then (2.63) holds.

2.1.6. Multiplier uniqueness for two-side inequality constraints. Consider the optimization problem (P_G) , in which \mathbb{F} is $\mathbb{R}^m = \mathbb{R}^{m_E} \times \mathbb{R}^{m_I}$ and

$$G = \{0_{\mathbb{R}^{m_E}}\} \times [l, u],$$

where for some vectors l and $u \in \overline{\mathbb{R}}^{m_I}$, $[l, u] := \{v \in \mathbb{R}^{m_I} : l \leq v \leq u\}$. Let $(x, \lambda) \in \mathbb{E} \times \mathbb{R}^m$ be a **stationary pair** of the problem (P_G) and denote by Λ the set (2.48) of **optimal multipliers** associated with x . Define the following index sets

$$\begin{aligned} I^l &:= \{i \in I : c_i(x) = l_i\}, & I^u &:= \{i \in I : c_i(x) = u_i\}, \\ I^{l0} &:= \{i \in I : c_i(x) = l_i, \lambda_i = 0\}, & I^{u0} &:= \{i \in I : c_i(x) = u_i, \lambda_i = 0\}, \\ I^{l-} &:= \{i \in I : c_i(x) = l_i, \lambda_i < 0\}, & I^{u+} &:= \{i \in I : c_i(x) = u_i, \lambda_i > 0\}. \end{aligned}$$

Show that the following properties are equivalent:

- (i) $\Lambda = \{\lambda\}$,
- (ii) there is no nonzero $\alpha \in \mathbb{R}^{|E \cup I^l \cup I^u|}$ such that

$$\sum_{i \in E \cup I^l \cup I^u} \alpha_i \nabla c_i(x) = 0, \quad \alpha_{I^l} \leq 0, \quad \text{and} \quad \alpha_{I^u} \geq 0,$$

- (iii) for any $v \in \mathbb{R}^{|E \cup I^l \cup I^u|}$, there is a $d \in \mathbb{E}$ such that

$$c'_{E \cup I^l \cup I^u}(x)d = v_{E \cup I^l \cup I^u}, \quad c'_{I^l}(x)d \leq v_{I^l}, \quad \text{and} \quad c'_{I^u}(x)d \geq v_{I^u},$$

- (iv) $c'_{E \cup I^l \cup I^u}(x)$ is surjective and there is a $d \in \mathbb{E}$ such that

$$c'_{E \cup I^l \cup I^u}(x)d = 0, \quad c'_{I^l}(x)d < 0, \quad \text{and} \quad c'_{I^u}(x)d > 0.$$

2.1.7. *Multiplier uniqueness for (P_G) with $G = \mathcal{S}_+^n$ [129].* Consider the problem (P_G) in which \mathbb{F} is the vector space \mathcal{S}^n formed of the symmetric matrices of order n equipped with the scalar product $\langle \cdot, \cdot \rangle : (A, B) \in \mathcal{S}^n \times \mathcal{S}^n \mapsto \langle A, B \rangle = \text{tr } AB \in \mathbb{R}$ (the trace of the matrix AB) and $G = \mathcal{S}_+^n$ is the cone of \mathcal{S}^n made of the positive semidefinite matrices. The problem reads

$$\begin{cases} \min f(x) \\ c(x) \in \mathcal{S}_+^n. \end{cases} \quad (2.65)$$

We also define $\mathcal{S}_-^n = -\mathcal{S}_+^n$. Let $(x, \lambda) \in \mathbb{E} \times \mathcal{S}^n$ be a stationary pair of (2.65). Let $r := \text{rank}(c(x))$ be the rank of the matrix $c(x) \in \mathcal{S}^n$, so that there is an $n \times (n - r)$ injective matrix V such that $\mathcal{N}(c(x)) = \mathcal{R}(V)$.

- 1) Show that λ is of the form $\lambda = V\Theta V^\top$, for some $\Theta \in \mathcal{S}_-^{n-r}$.
- 2) Show that $\text{rank}(\lambda) + \text{rank}(c(x)) \leq n$.

From now on, we assume that *strict complementarity* holds, which means that $\text{rank}(\lambda) + \text{rank}(c(x)) = n$.

- 3) Show that $\mathbf{T}_\lambda^f(\mathcal{N}_{c(x)} \mathcal{S}_+^n) = \mathbf{T}_\lambda(\mathcal{N}_{c(x)} \mathcal{S}_+^n) = \{V\Theta V^\top : \Theta \in \mathcal{S}^{n-r}\}$.
- 4) Show that x has a *unique* associated multiplier λ if and only if

$$c'(x)\mathbb{E} + \{Z \in \mathcal{S}^n : V^\top ZV = 0\} = \mathcal{S}^n.$$

2.1.8. *Optimality conditions of the first order for a problem with an additional set-inclusion constraint.* Prove propositions 2.26, 2.27, 2.28, 2.29 and 2.30.

2.1.9. *Optimality conditions of the first order for a composite problem.* Prove propositions 2.32, 2.33 and 2.34.

2.2 Second Order Optimality Conditions for (P_{EI})

Second order optimality conditions are used to determine whether a stationary point is a local minimizer or maximizer; sometimes these conditions are not precise enough to conclude. One makes a distinction between *necessary* optimality conditions (those that are implied by a local minimizer) and *sufficient* optimality conditions (those that guarantee that a given point is a local minimizer).

The necessary optimality conditions of the second order for the problem with equality and inequality constraints (P_{EI}) defined in section 1.4.4 can neither be obtained as easily nor be written as simply as for a problem with only equality constraints (theorem 1.33). There is a common feature however, which is that it is the Hessian of the Lagrangian that intervenes in these conditions; the reason is that, like for the equality constrained problem, the gradient of the Lagrangian vanishes at a local solution to the problem (compare theorems 1.30 and 1.40). But there are two main differences. First, it is not on the tangent cone to the feasible set that the Hessian of the Lagrangian is positive semi-definite, but on a smaller one, called the *critical cone* (this is further discussed in section 2.2.1). The second difference comes from the choice of the optimal multiplier that intervenes in the Hessian of the Lagrangian (for an inequality constrained problem, it is frequent that the set of optimal multipliers associated with a given solution is not reduced to a singleton): the key observation is that the optimal multiplier must be chosen according to the given critical direction (this is further discussed in section 2.2.2).

Getting second order *sufficient* conditions of optimality is not a difficult task (section 2.2.4), but getting the second order *necessary* conditions of optimality is much more serious (section 2.2.3). There are many possibilities to establish them. Our strategy is the following. The approach assumes at once that the Mangasarian-Fromovitz constraint qualification condition (CQ-MF) holds at the solution. This one is strong enough to show the existence of paths in the feasible set, emanating from the considered solution. Then, the behavior of the objective of the problem is examined along these paths, which allows us to derive *weak* necessary optimality conditions of the second order. These optimality conditions can be reinforced in the presence of stronger constraint qualification conditions. It is the subject of exercise ?? to consider the case of (CQ-A) and (CQ-LI) and to ask to show that a *strong* form of the second order optimality conditions are then obtained (this is not surprising for (CQ-LI), since then the optimal multiplier is uniquely determined).

2.2.1 Critical Cone

It is tempting to try to generalize the necessary optimality conditions of the second order of problem (P_E) , stated in theorem 1.33, to problem (P_{EI}) . This extension could be that, at a stationary pair (x_*, λ_*) , one must have $\langle L_* d, d \rangle \geq 0$ for all tangent directions $d \in T_{x_*} X_{EI}$ (we have denoted by $L_* := \nabla_{xx}^2 \ell(x_*, \lambda_*)$ the Hessian of the Lagrangian at the considered stationary pair). This result is not correct, since the tangent cone $T_{x_*} X_{EI}$ is not the appropriate one, as shown by the following example

$$\min \left\{ \frac{-1}{x+1} : x \in \mathbb{R}_+ \right\}. \quad \begin{array}{c} \mathbb{R}_+ \\ \hline f(x) = -1/(x+1) \end{array} \quad (2.66)$$

This problem has for unique stationary point $(x_*, \lambda_*) = (0, 1)$ and the tangent cone at x_* reads $T_{x_*} X_{EI} = \mathbb{R}_+$, so that one can take $d = 1$ as tangent direction, but $\langle L_* d, d \rangle = -2$ has not the right sign. We shall see that $\langle L_* d, d \rangle \geq 0$, but for directions d in a cone that is *smaller* than the tangent cone.

Looking for a smaller cone, one could imagine it as a tangent cone to a smaller set than X_{EI} . Observe now that a solution x_* to (P_{EI}) also minimises f locally on the smaller set

$$X_{EI}^{\bar{\bar{}}} := \{x \in \mathbb{E} : c_{E \cup I_*^0}(x) = 0, c_{I \setminus I_*^0}(x) < 0\}.$$

Theorem 1.33 tells us that $\langle L_* d, d \rangle \geq 0$ for all directions $d \in T_{x_*} X_{EI}^{\bar{\bar{}}}$ and all associated multipliers $\lambda_* \in \Lambda_*$ (these are also multipliers of the problem consisting in minimizing f on $X_{EI}^{\bar{\bar{}}}$). We shall see, however, that $T_{x_*} X_{EI}^{\bar{\bar{}}}$ is too small, in the sense that it does not allow us to establish *sufficient* optimality condition of the second order. Consider indeed the problem

$$\min \{-x^2 : x \in \mathbb{R}_+\}. \quad \begin{array}{c} \mathbb{R}_+ \\ \hline f(x) = -x^2 \end{array} \quad (2.67)$$

The point $x_* = 0$ is a stationary point of this problem (for arbitrary multipliers in \mathbb{R}_+). Since $X_{EI}^{\bar{\bar{}}} = \{0\}$, it follows that $T_{x_*} X_{EI}^{\bar{\bar{}}} = \{0\}$. Furthermore, the Hessian of

the Lagrangian at (x_*, λ_*) is $L_* = -2$ (for any chosen multiplier). Now, $\langle L_* d, d \rangle$ is positive for all d in $T_{x_*} X_{EI}^- \setminus \{0\} = \emptyset$, but x_* not a local minimum of the problem.

The appropriate cone will turn out to be the **linearizing cone** $T'_{x_*}(\text{Sol}(P_{EI}))$ to the solution set of (P_{EI}) , written as follows

$$\text{Sol}(P_{EI}) := \{x \in X_{EI} : f(x) \leq f(x_*)\}. \quad (2.68)$$

This cone is smaller than the **linearizing cone** to the feasible set X_{EI} at x_* , but sufficiently large to yield sufficient optimality condition of the second order (theorem 2.39). It is called the *critical cone* of the problem.

Definition 2.35 The *critical cone* of problem (P_{EI}) at a feasible point $x \in X_{EI}$ is the polyhedral cone denoted and defined by

$$C(x) := \{d \in \mathbb{E} : c'_E(x) \cdot d = 0, c'_{I^0(x)}(x) \cdot d \leq 0, f'(x) \cdot d \leq 0\}. \quad (2.69a)$$

A direction $d \in C(x)$ is called a *critical direction* at x . We shall use the simplified notation $C_* := C(x_*)$. \square

Therefore, the critical cone is the **linearizing cone** (1.56), with the additional constraint $f'(x) \cdot d \leq 0$ on its directions d .

In the example (2.66), $C_* = \{0\}$ is smaller than the tangent cone $T_{x_*} X_{EI} = \mathbb{R}_+$. In the example (2.67), $C_* = \mathbb{R}_+$ is larger than the tangent cone $T_{x_*} X_{EI}^- = \{0\}$. It is remarkable that the optimality at the second order can be synthetised by means of the single critical cone, while the two previous problems cover very different situations.

At a stationary pair (x_*, λ_*) , the critical cone at x_* also reads

$$C_* = \{d \in \mathbb{E} : c'_E(x_*) \cdot d = 0, c'_{I^0_*}(x_*) \cdot d \leq 0, f'(x_*) \cdot d = 0\}, \quad (2.69b)$$

$$= \{d \in \mathbb{E} : c'_{E \cup I^{0+}_*}(x_*) \cdot d = 0, c'_{I^{00}_*}(x_*) \cdot d \leq 0\}, \quad (2.69c)$$

where we have used the index sets

$$\begin{aligned} I_*^0 &:= \{i \in I : c_i(x_*) = 0\}, \\ I_*^{0+} &:= \{i \in I_*^0 : (\lambda_*)_i > 0\}, \\ I_*^{00} &:= \{i \in I_*^0 : (\lambda_*)_i = 0\}. \end{aligned}$$

The expressions (2.69b) and (2.69c) can be obtained by using the optimality conditions (1.58). Observe finally that, if strict complementarity holds, in the sense (1.59), the critical cone simply reads

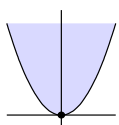
$$C_* = \{d \in \mathbb{E} : c'_{E \cup I^0_*}(x_*) \cdot d = 0\}, \quad (2.69d)$$

which is the **linearizing cone** $T'_{x_*} X_{EI}^-$ (it is a linear subspace in \mathbb{E}). Without strict complementarity, the linear subspace (2.69d) is included in C_* , itself included in $T'_{x_*} X_{EI}$.

2.2.2 Three Instructive Examples

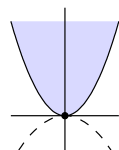
Another difficulty in establishing second order optimality conditions for problem (P_{EI}) comes from the fact that the optimal multiplier intervening in the Hessian of the Lagrangian $L_* := \nabla_{xx}^2 \ell(x_*, \lambda_*)$ must be chosen according to the critical direction. In others words, the proof of theorem 1.33, based on the development of the Lagrangian $\ell(\cdot, \lambda_*)$ for an a priori given multiplier λ_* no longer works for a solution satisfying only the Mangasarian-Fromovitz constraint qualification **(CQ-MF)** (see exercise ?? for the constraint qualification **(CQ-A)** and **(CQ-LI)**, for which such a technique can be used). We give three examples to better understand the situation and to learn how to select the correct sequence of quantifiers to apply to $d \in C_*$ and $\lambda_* \in \Lambda_*$.

Consider first the simple two variable problem

$$\left\{ \begin{array}{l} \min x_2 \\ x_2 \geq x_1^2. \end{array} \right. \quad \begin{array}{c} \text{---} \\ \text{---} \\ | \\ \text{---} \\ \text{---} \\ \text{---} \end{array} \quad (2.70)$$


Its feasible set is represented in the right-hand side above. The problem has for unique solution $x_* = (0, 0)$ and there is a unique associated multiplier $\lambda_* = 1$. Since the constraint is active at x_* , (x_*, λ_*) is also a stationary pair of the equality constrained problem $\min\{x_2 : x_2 = x_1^2\}$, so that the Hessian of the Lagrangian $L_* := \nabla_{xx}^2 \ell(x_*, \lambda_*) = \text{diag}(2, 0)$ must be positive semi-definite on the tangent space $\{d \in \mathbb{R}^2 : d_2 = 0\}$. This is the most simple situation that can occur. Below, we shall say that the *strong second order optimality conditions* hold, meaning that, for any optimal multiplier λ_* (there is a single one here), L_* is positive semi-definite on the critical cone. These conditions are verified if there is a unique multiplier, like here, or when the constraint qualification conditions **(CQ-A)** or **(CQ-LI)** hold (see exercise ??).

Consider now a variation of problem (2.70), in which a superfluous constraint is added:

$$\left\{ \begin{array}{l} \min x_2 \\ x_2 \geq x_1^2 \\ x_2 \geq -\frac{1}{2}x_1^2. \end{array} \right. \quad \begin{array}{c} \text{---} \\ \text{---} \\ | \\ \text{---} \\ \text{---} \\ \text{---} \end{array} \quad (2.71)$$


The second constraint does not modify the solution to the problem, which is again $x_* = (0, 0)$, but there are now several optimal multipliers, forming the set $\Lambda_* = \{\lambda \in \mathbb{R}_+^2 : \lambda_1 + \lambda_2 = 1\}$. By taking the multiplier $\lambda_* = (1, 0)$, a vertex of Λ_* , one ignores the second constraint, which is appropriate, and one has the preceding result on the positive semi-definiteness of $L_* = \text{diag}(2, 0)$ on the critical cone $C_* = \{d \in \mathbb{R}^2 : d_2 = 0\}$. In contrast, with $\lambda_* = (0, 1)$, the other vertex of Λ_* , the Hessian of the Lagrangian $L_* = \text{diag}(-1, 0)$ is *negative* definite on C_* . This is normal; with that λ_* , the Lagrangian $\ell(\cdot, \lambda_*)$ can only see the second constraint, hence ignoring the first one, and $(0, 0)$ is only a stationary point of the problem $\min\{x_2 : x_2 \geq -\frac{1}{2}x_1^2\}$, not a local minimum. Below, we shall say that the *semi-strong second order optimality conditions* hold, meaning that, there exists an optimal multiplier λ_* , such that L_* is positive semi-definite on the critical cone.

An inequality constrained optimization problem is not always as simple as problem (2.71), in which one can locally (around the solution) discard all the constraints but

one, while keeping optimality of the solution. Sometimes, each time a constraint is discarded, the optimality is lost and this phenomenon is reflected in the second order optimality conditions. Here is an example with three variables:

$$\begin{cases} \min x_3 \\ x_3 \geq (x_1 + x_2)(x_1 - x_2) \\ x_3 \geq (x_2 + 3x_1)(2x_2 - x_1) \\ x_3 \geq (2x_2 + x_1)(x_2 - 3x_1). \end{cases} \quad \begin{array}{ccc} \begin{array}{c} \text{---} \\ | \\ \text{---} \\ | \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ | \\ \text{---} \\ | \\ \text{---} \end{array} & \begin{array}{c} \text{---} \\ | \\ \text{---} \\ | \\ \text{---} \end{array} \end{array} \quad (2.72)$$

The three pictures in the right-hand side above represent, for each of the three constraints, the coordinates (x_1, x_2) giving a positive value of their right-hand side. We see that, for any nonzero (x_1, x_2) , one of these right-hand side is positive. As a result, the unique solution to the problem is $x_* = 0$. Furthermore, the set of associated optimal multipliers is the unit simplex $\Lambda_* = \{\lambda \in \mathbb{R}_+^3 : \lambda_1 + \lambda_2 + \lambda_3 = 1\}$. Finally, the Hessian of the Lagrangian reads

$$L(x, \lambda) = \begin{pmatrix} 2\lambda_1 - 6(\lambda_2 + \lambda_3) & 5(\lambda_2 - \lambda_3) & 0 \\ 5(\lambda_2 - \lambda_3) & -2\lambda_1 + 4(\lambda_2 + \lambda_3) & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Whatever is the vector λ_* chosen in Λ_* , L_* is not positive semi-definite on the critical cone $C_* = \{d \in \mathbb{R}^3 : d_3 = 0\}$. Indeed, the element $(1, 1)$ of L_* has the value $8\lambda_1 - 6$ and the element $(2, 2)$ is $4 - 6\lambda_1$, so that the positive definiteness of L_* would require to have $\lambda_1 > 3/4$ and $\lambda_1 < 2/3$, an impossibility. Below, we shall say that the *weak second order optimality conditions* hold, meaning that, for all critical direction d , there is a multiplier λ_* (depending on d), such that $\langle L_* d, d \rangle \geq 0$.

2.2.3 Second Order Necessary Optimality Conditions

Theorem 2.37 below states necessary optimality conditions of the second order, assuming that the Mangasarian-Fromovitz constraint qualification (CQ-MF) holds at the considered solution. This assumption is not very restrictive, since it is the weakest of the sufficient constraint qualification conditions presented in section 1.4.4. Theorem 2.39 states sufficient optimality conditions of the second order, without constraint qualification conditions.

The necessary conditions lie on the examination of the behavior of the criterion f along paths $t \mapsto \xi(t)$ emanating from the considered solution x_* at $t = 0$ and remaining in the feasible set X_{EI} for small *positive* values of the parameter t . More paths of this type are considered, more information will be obtained. The approach is the same for the equality constrained problem (P_E) , but it turned out that all the paths having the same tangent at the origin gave the same second order information for that problem. In the presence of inequality constraints, for a given tangent direction $d \in T_{x_*} X_{EI}$, it is useful to distinguish paths having $d = \xi'(0)$ as tangent at the origin, but having distinct curvatures $h = \xi''(0)$. Therefore, for given d and h , one looks for paths $t \mapsto \xi(t)$ such that

$$\xi(0) = x_*, \quad \xi'(0) = d, \quad \xi''(0) = h, \quad \xi(t) \in X_{EI}, \text{ for small } t \geq 0. \quad (2.73)$$

Tangency and curvature must satisfy compatibility relations so that (2.73) holds. Let us guess what can be these conditions; the first three (2.74a)-(2.74c) determined below

are necessary and the last one (2.74d) has a sufficient flavor. Lemma 2.36 will show that they can be satisfied under (CQ-MF) and that, all together, they are indeed sufficient to guarantee the existence of a path satisfying (2.73).

Consider first the equality constraints: the map ξ must verify $c_E(\xi(t)) = 0$, for all small $t \geq 0$. Differentiating once the vanishing map $t \mapsto c_E(\xi(t))$, we get $c'_E(\xi(t)) \cdot \xi'(t) = 0$ for all small $t \geq 0$. At $t = 0$, (2.73) gives

$$c'_E(x_*) \cdot d = 0. \quad (2.74a)$$

Differentiating once the vanishing map $t \mapsto c'_E(\xi(t)) \cdot \xi'(t)$, we get $c''_E(\xi(t)) \cdot (\xi'(t))^2 + c'_E(\xi(t)) \cdot \xi''(t) = 0$ for all small $t \geq 0$. At $t = 0$, (2.73) gives

$$c''_E(x_*) \cdot d^2 + c'_E(x_*) \cdot h = 0. \quad (2.74b)$$

Consider now the inequality constraints: the map ξ must verify $c_I(\xi(t)) \leq 0$, for all small $t \geq 0$. If $i \in I \setminus I^0(x_*)$, $c_i(x_*) < 0$, so that $c_i(\xi(t)) \leq 0$ for small $t \geq 0$ if both c_i and ξ are continuous. Consider now the indices $i \in I^0(x_*)$ of the active constraints at x_* . Taking a first order development of the map $t \mapsto c_i(\xi(t))$ around $t = 0$ and using (2.73), we get $c_i(x_*) + [c'_i(x_*) \cdot d] t + o(t) \leq 0$ for all small $t \geq 0$. Since $c_i(x_*) = 0$, dividing by $t > 0$ and taking the limit when $t \downarrow 0$, we see that we must have

$$c'_{I^0}(x_*) \cdot d \leq 0. \quad (2.74c)$$

Taking a second order development of $t \mapsto c_i(\xi(t))$, we see that we must have

$$\underbrace{c_i(x_*)}_{=0} + t \underbrace{c'_i(x_*) \cdot d}_{\leq 0} + \frac{t^2}{2} (c''_i(x_*) \cdot d^2 + c'_i(x_*) \cdot h) + o(t^2) \leq 0.$$

Here, one cannot deduce a necessary condition on d and h , but to have satisfaction of the previous inequality it is *sufficient* to impose

$$c''_i(x_*) \cdot d^2 + c'_i(x_*) \cdot h \leq -\varepsilon e, \quad (2.74d)$$

where $\varepsilon > 0$ and e is the vector of all ones.

Note that d satisfying (2.74a) and (2.74c) is, by definition, a direction of the **linearizing cone** to X_{EI} at x_* , hence a tangent direction to X_{EI} at x_* if the constraints are qualified at x_* .

Lemma 2.36 (existence of a feasible path) *Let $x_* \in X_{EI}$. Suppose that c_E is \mathcal{C}^2 in a neighborhood of x_* , that c_{I^0} is twice differentiable at x_* , and that c_{I^*} is continuous at x_* . Suppose also that the Mangasarian-Fromovitz constraint qualification condition (CQ-MF) holds at x_* . Let $\varepsilon > 0$. Then,*

- 1) *for all $d \in \mathbb{T}_{x_*} X_{EI}$, there exists $h \in \mathbb{E}$ such that (2.74b) and (2.74d) hold,*
- 2) *for all $(d, h) \in \mathbb{E} \times \mathbb{E}$ satisfying (2.74), there exists a path $\xi : t \in \mathbb{R} \mapsto \xi(t) \in \mathbb{E}$ of class \mathcal{C}^2 , defined for $|t|$ sufficiently small, such that (2.73) holds.*

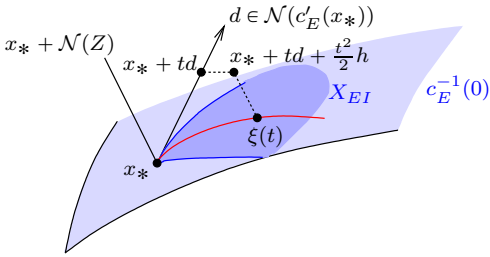
Proof. 1) This is a direct consequence of (CQ-MF), expressed by the point (ii) of proposition 1.39.

2) Let $A_E := c'_E(x_*)$ be the Jacobian of c_E at x_* . This one being surjective, one can find a linear operator $Z : \mathbb{E} \rightarrow \mathbb{R}^{n-m_E}$, such that

$$\begin{pmatrix} A_E \\ Z \end{pmatrix} : \mathbb{E} \rightarrow \mathbb{R}^n \text{ is bijective.} \quad (2.75)$$

Note that this property implies that $\mathcal{N}(A_E) \cap \mathcal{N}(Z) = \emptyset$ and $\mathcal{N}(A_E) + \mathcal{N}(Z) = \mathbb{E}$, so that \mathbb{E} can be written as a direct sum: $\mathbb{E} = \mathcal{N}(A_E) \oplus \mathcal{N}(Z)$.

For $d \in \mathbb{T}_{x_*} X_{EI}$ and $h \in \mathbb{E}$ given by point 1, one considers the function $F : \mathbb{E} \times \mathbb{R} \rightarrow \mathbb{R}^n$ defined at $(\xi, t) \in \mathbb{E} \times \mathbb{R}$ by

$$F(\xi, t) = \begin{pmatrix} c_E(\xi) \\ Z(\xi - x_* - td - \frac{t^2}{2}h) \end{pmatrix}.$$


If $F(\xi, t) = 0$, $\xi \in \mathbb{E}$ is necessarily a point on the manifold $c_E^{-1}(0)$, while $\xi - x_* - td - (t^2/2)h$ is a displacement in the null space of Z , which is complementary to the null space of A_E , itself tangent to the manifold $c_E^{-1}(0)$ at x_* . The path $t \mapsto \xi(t)$ is obtained by imposing the nullity of $F(\xi, t)$ and using the implicit function theorem. By assumption, the function F is of class \mathcal{C}^2 in a neighborhood of $(x_*, 0)$, $F(x_*, 0) = 0$ and $F'_\xi(x_*, 0)$ is nonsingular. By the implicit function theorem, there exists a function $t \mapsto \xi(t)$, defined for $|t|$ small, of class \mathcal{C}^2 , such that $F(\xi(t), t) = 0$ for all small $|t|$ and $\xi(0) = x_*$. By differentiating once $F(\xi(t), t) = 0$ at $t = 0$, one gets

$$\begin{pmatrix} A_E \\ Z \end{pmatrix} \xi'(0) = \begin{pmatrix} 0 \\ Zd \end{pmatrix} = \begin{pmatrix} A_E \\ Z \end{pmatrix} d,$$

because $A_E d = 0$. Therefore, (2.75) shows that $\xi'(0) = d$. By differentiating twice $F(\xi(t), t) = 0$ at $t = 0$, one gets

$$\begin{pmatrix} A_E \\ Z \end{pmatrix} \xi''(0) = \begin{pmatrix} -c''_E(x_*) \cdot d^2 \\ Zh \end{pmatrix} = \begin{pmatrix} A_E \\ Z \end{pmatrix} h,$$

by (2.74b). Therefore, $\xi''(0) = h$. Finally, $\xi(t) \in X_{EI}$ for small $t \geq 0$, thanks to the choice of h verifying (2.74c) and (2.74d). \square

Theorem 2.37 (NC2 for (P_{EI}) with $(CQ-MF)$) Suppose that x_* is a local solution to (P_{EI}) , that f and c_E are \mathcal{C}^2 in a neighborhood of x_* , that c_{I_0} is twice differentiable at x_* , that c_{I^*} is continuous at x_* , and that the Mangasarian-Fromovitz constraint qualification condition $(CQ-MF)$ holds at x_* . Then

$$\forall d \in C_* : \max_{\lambda_* \in \Lambda_*} \langle L_* d, d \rangle \geq 0, \quad (2.76)$$

where $L_* := \nabla_{xx}^2 \ell(x_*, \lambda_*)$ is the Hessian of the Lagrangian.

Proof. Let $d \in C_* \subseteq T'_{x_*} X_{EI}$ be fixed. For each choice of $h \in \mathbb{E}$ verifying (2.74b) and (2.74d), lemma 2.36 ensures the existence of a path $t \mapsto \xi(t)$ of class \mathcal{C}^2 satisfying (2.73). The direction h will be chosen to get the best information. Let us see this. A second order development of $t \mapsto f(\xi(t))$ at $t = 0$ reads

$$f(\xi(t)) = f(x_*) + t f'(x_*) \cdot d + \frac{t^2}{2} [f''(x_*) \cdot d^2 + f'(x_*) \cdot h] + o(t^2).$$

By optimality, $f(\xi(t)) \geq f(x_*)$, for $t \geq 0$ small (since then $\xi(t) \in X_{EI}$), and $f'(x_*) \cdot d \leq 0$ (since $d \in C_*$). Therefore,

$$0 \leq \frac{t^2}{2} [f''(x_*) \cdot d^2 + f'(x_*) \cdot h] + o(t^2).$$

Dividing by $t^2 > 0$ and taking the limit when $t \downarrow 0$, we get

$$0 \leq f''(x_*) \cdot d^2 + f'(x_*) \cdot h. \quad (2.77)$$

To get as much information as possible, it is better to find an h minimizing the right-hand side of the above inequality. This remark leads us to the following linear problem in $h \in \mathbb{E}$:

$$\begin{cases} \inf f'(x_*) \cdot h + f''(x_*) \cdot d^2 \\ c'_i(x_*) \cdot h + c''_i(x_*) \cdot d^2 = 0, & \text{for } i \in E \\ c'_i(x_*) \cdot h + c''_i(x_*) \cdot d^2 + \epsilon \leq 0, & \text{for } i \in I_*^0. \end{cases} \quad (2.78)$$

This problem is feasible (i.e., its feasible set is nonempty, thanks to (CQ-MF)) and bounded (by zero). Therefore, by (1.63), the problem has a solution and by the sentence after (1.66), there is no duality gap between its optimal value and the optimal value of its dual: these are both equal and nonnegative by (2.77). The dual problem is the following linear optimization problem in $\lambda \in \mathbb{R}^m$

$$\begin{cases} \sup \langle L(x_*, \lambda) d, d \rangle + \epsilon \|\lambda_I\|_1 \\ \nabla_x \ell(x_*, \lambda) = 0 \\ \lambda_{I_*^0} \geq 0 \\ \lambda_{I_*^c} = 0. \end{cases}$$

Observe now that the feasible set of this last problem is the set of optimal multipliers Λ_* . Therefore, we have shown that

$$0 \leq \max_{\lambda_* \in \Lambda_*} \langle L_* d, d \rangle + \epsilon \|(\lambda_*)_I\|_1.$$

We have used the operator “max”, since Λ_* is compact by (CQ-MF). Since $\epsilon > 0$ is arbitrary and since Λ_* is compact, one gets the result by taking $\epsilon \rightarrow 0$. \square

When (CQ-MF) holds at x_* , A_* is compact, so that the condition (2.76) also reads

$$\forall d \in C_*, \exists \lambda_* \in A_* : \langle L_* d, d \rangle \geq 0. \quad (2.79a)$$

It is then said that the *weak necessary optimality conditions of the second order* hold. They are satisfied under the assumptions given by theorem 2.37. This is the case in the example 2.72. If stronger conditions holds, namely

$$\exists \lambda_* \in A_*, \forall d \in C_* : \langle L_* d, d \rangle \geq 0, \quad (2.79b)$$

it is said that the *semi-strong necessary optimality conditions of the second order* hold. They are satisfied in example 2.71. If even stronger conditions holds, namely

$$\forall \lambda_* \in A_*, \forall d \in C_* : \langle L_* d, d \rangle \geq 0, \quad (2.79c)$$

it is said that the *strong necessary optimality conditions of the second order* hold. They are satisfied in example 2.70.

The strong second order necessary conditions of optimality (2.79c) are clearly satisfied when there is a unique optimal multiplier, since then (2.79a), (2.79b), and (2.79c) contain the same information and (2.79a) holds by theorem 2.37. These conditions are also satisfied when some constraint qualification conditions stronger than (CQ-MF) are satisfied. The next theorem claims that this is the case when (CQ-A) or (CQ-LI) holds; this result was the one presented in most textbooks on nonlinear optimization prior to 1980, say; see [53, 95, 34, 8; 1968-1976] for instance.

Theorem 2.38 (NC2 for (P_{EI}) with (CQ-A) or (CQ-LI)) *Suppose that x_* is a local solution to (P_{EI}) , that f and c are twice differentiable at x_* , and that the constraint qualification conditions (CQ-A) or (CQ-LI) hold at x_* . Then, the strong second order necessary conditions of optimality (2.79c) hold.*

Proof. See exercise ??.

□

The numerical verification of the necessary optimality conditions of the second order is not an easy task. Even when the semi-strong conditions (2.79b) hold for an optimal multiplier λ_* , one has to verify that the quadratic form $d \mapsto \langle L_* d, d \rangle$ associated with the Hessian of the Lagrangian is positive semi-definite on the critical cone C_* , which is polyhedral; in other words, L_* is C_* -copositive [70, 16, 68]. In all generality, such a verification is an NP-hard problem [104, 43]. Now, if **strict complementarity** also holds, the critical cone reduces to the linear subspace (2.69d) and the verification of the positive semi-definiteness of $d \mapsto \langle L_* d, d \rangle$ on this subspace is then a simple linear algebra operation.

2.2.4 Second Order Sufficient Optimality Conditions

The next proposition gives sufficient conditions of optimality of the second order for problem (P_{EI}) . It is worth noting that these do not call on a constraint qualification assumption. The fact that the critical cone also intervenes in these conditions is an

evidence of its relevance. The inequality (2.81) is known as the *quadratic growth property*. It tells us that f grows at least quadratically on the “interior” of the feasible set X_{EI} .

Theorem 2.39 (SC2 for (P_{EI})) *Suppose that f and $c_{E \cup I^0}$ are differentiable on a neighborhood of a point $x_* \in \mathbb{E}$ and twice differentiable at x_* . Suppose also that the set Λ_* of optimal multipliers λ_* such that (x_*, λ_*) verifies the KKT optimality conditions (1.58) is nonempty. Suppose finally that the following equivalently properties hold ($\|\cdot\|$ is an arbitrary norm)*

$$\forall d \in C_* \setminus \{0\}, \exists \lambda_* \in \Lambda_* : \langle L_* d, d \rangle > 0, \quad (2.80a)$$

$$\exists \bar{\gamma} > 0, \forall d \in C_*, \exists \lambda_* \in \Lambda_* : \langle L_* d, d \rangle \geq \bar{\gamma} \|d\|^2. \quad (2.80b)$$

Then, for all $\gamma \in [0, \bar{\gamma})$, there exists a neighborhood V of x_ such that*

$$\forall x \in (X_{EI} \cap V) \setminus \{x_*\} : f(x) > f(x_*) + \frac{\gamma}{2} \|x - x_*\|^2. \quad (2.81)$$

In particular, x_ is a strict local minimum of (P_{EI}) .*

Proof. Let us first show that (2.80a) and (2.80b) are equivalent. It is clear that (2.80b) implies (2.80a). Let us show the contrapositive of the reverse implication, assuming that (2.80b) does not hold. Then, there would exist a sequence $\{d_k\} \subseteq C_*$ such that

$$\|d_k\| = 1 \quad \text{and} \quad \langle L(x_*, \lambda_*) d_k, d_k \rangle \rightarrow 0,$$

where λ_* is arbitrary fixed in Λ_* . Since C_* is closed, one could extract a converging subsequence $d_k \rightarrow d \in C_* \setminus \{0\}$. Then, for any $\lambda_* \in \Lambda_*$, we get $\langle L_* d, d \rangle = 0$, for some $d \in C_* \setminus \{0\}$. This contradicts (2.80a).

We prove the main claim of the theorem by contradiction, assuming that one can find $\gamma \in [0, \bar{\gamma})$ and a sequence $\{x_k\} \subseteq X_{EI}$ such that $x_k \rightarrow x_*$, $x_k \neq x_*$, and

$$f(x_k) \leq f(x_*) + \frac{\gamma}{2} \|x_k - x_*\|^2. \quad (2.82)$$

Extracting a subsequence if needed, one can assume that with $t_k := \|x_k - x_*\|$, there holds

$$\frac{x_k - x_*}{t_k} \rightarrow d.$$

Therefore, $d \in T_{x_*} X_{EI} \setminus \{0\}$. Furthermore, from (2.82) and $f(x_k) = f(x_*) + f'(x_*) \cdot (x_k - x_*) + o(\|x_k - x_*\|)$ (differentiability of f at x_* , one gets $f'(x_*) \cdot d \leq 0$. We have shown that $d \in C_* \setminus \{0\}$.

To get a contradiction, we take a second order expansion of the Lagrangian $\ell(\cdot, \lambda_*)$, where λ_* is the multiplier associated with d by (2.80b):

$$\ell(x_k, \lambda_*) = \ell(x_*, \lambda_*) + \frac{1}{2} \ell''_{xx}(x_*, \lambda_*) \cdot (x_k - x_*)^2 + o(\|x_k - x_*\|^2).$$

By the stationarity of x_* (see (1.58)), it follows that $\ell(x_*, \lambda_*) = f(x_*)$. By the feasibility of x_k and $(\lambda_*)_I \geq 0$, one gets $\ell(x_k, \lambda_*) \leq f(x_k)$, which does not exceed $f(x_*) + \frac{\gamma}{2} \|x_k - x_*\|^2$ by (2.82). Therefore,

$$\frac{\gamma}{2}\|x_k - x_*\|^2 \geq \frac{1}{2}\langle L_*(x_k - x_*), x_k - x_* \rangle + o(\|x_k - x_*\|^2).$$

Dividing by t_k^2 and taking the limit yield

$$\langle L_*d, d \rangle \leq \gamma\|d\|^2,$$

which contradicts (2.80b), since $\gamma < \bar{\gamma}$ and $d \in C_* \setminus \{0\}$. \square

Condition (2.81) is known as the *quadratic growth property* of f in X_{EI} around x_* .

The equivalent conditions (2.80a) and (2.80b) are called the *weak sufficient optimality conditions of the second order*. Obviously, the conclusion of the theorem is still true if in (2.80a), λ_* can be taken independently of d :

$$\exists \lambda_* \in \Lambda_*, \forall d \in C_* \setminus \{0\} : \langle L_*d, d \rangle > 0. \quad (2.83)$$

It is then said that *semi-strong sufficient optimality conditions of the second order* hold. Obviously also, the conclusion of the theorem remains true, if λ_* can be chosen arbitrarily:

$$\forall \lambda_* \in \Lambda_*, \forall d \in C_* \setminus \{0\} : \langle L_*d, d \rangle > 0. \quad (2.84)$$

It is then said that *strong sufficient optimality conditions of the second order* hold.

A natural question to ask is whether the conditions of the previous theorem not only ensure that x_* is a strict local minimum, but that it is also an *isolated minimum*. The answer is negative, as shown by the following counter-example [121; (2.5)].

Counter-example 2.40 (strict but nonisolated minimum) Consider the following problem in $x \in \mathbb{R}$:

$$\min \left\{ \frac{1}{2}x^2 : c(x) = 0 \right\}, \quad \text{where} \quad c(x) = \begin{cases} x^6 \sin \frac{1}{x} & \text{if } x \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The conditions of theorem 2.39 are satisfied at the solution $x_* = 0$ with the multiplier $\lambda_* = 0$, since $\ell''_{xx}(x_*, \lambda_*) = 1$ but x_* is not an isolated minimum. Indeed, 0 is an accumulation point of the feasible set, which reads $\{0\} \cup \{k\pi : k \in \mathbb{N}\}$, and every feasible point is a local minimizer. \square

The undesirable situation of counter-example 2.40 would not occur if the (CQ-MF) constraint qualification was added to the assumptions of the theorem; see [121; § 2].

Notes

There are many ways to get second order conditions of optimality. The one followed here, which is based on the design of paths in the feasible set, on the expression of the behavior of the objective along these path, and on the dualization of this expression is the one followed in the short account of Gauvin [60].

Second order optimality conditions for (P_{EI}) were given in [53; 1968] under restrictive conditions. The conditions related here can be found in [72, 13, 15; 1979-1982]. See also the exact penalty viewpoint of Burke [28].

Exercises

2.2.1. *Second order optimality conditions.* Consider the following nonlinear optimization problem in $x \in \mathbb{R}^2$:

$$\begin{cases} \min -\frac{1}{2}(x_1^2 + x_2^2) \\ x_2 \geq x_1^2 - 1 \\ x_1 \geq 0. \end{cases}$$

Using the Lagrangian $\ell : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}$ defined at (x, λ) by

$$\ell(x, \lambda) = -\frac{1}{2}(x_1^2 + x_2^2) + \lambda_1(x_1^2 - x_2 - 1) + \lambda_2(-x_1),$$

it can be shown that the first order optimality conditions are verified by the following primal-dual pairs

$$x = 0 \quad \text{and} \quad \lambda = 0, \tag{2.85a}$$

$$x = (0, -1) \quad \text{and} \quad \lambda = (1, 0), \tag{2.85b}$$

$$x = (\sqrt{2}/2, -1/2) \quad \text{and} \quad \lambda = (1/2, 0). \tag{2.85c}$$

Using the second order optimality conditions, determine analytically which of the points in (2.85) are (strict) local minimum, (strict) local maximum, or undetermined.

3 Perturbation Analysis

3.1 Linear System ▲

3.2 Nonlinear System ▲

3.3 Optimization ▲

Let \mathbb{E} and \mathbb{F} be a vector spaces, Ω be an *open convex set* of \mathbb{E} , and \mathcal{P} be a topological space. Consider the family of optimization problems

$$(P_K^p) \quad \begin{cases} \min_x f(x, p) \\ c(x, p) \in K, \end{cases}$$

parametrized by $p \in \mathcal{P}$, in which $f : \Omega \times \mathcal{P} \rightarrow \mathbb{R}$ and $c : \Omega \times \mathcal{P} \rightarrow \mathbb{F}$ are smooth functions, and $K \subseteq \mathbb{F}$ is a nonempty *closed convex cone*. This problem will be viewed as a perturbation of a problem of the form (2.1), that is supposed to be recovered when p is set to some reference parameter $p_0 \in \mathcal{P}$, in the sense that $f(\cdot, p_0) = f(\cdot)$ and $c(\cdot, p_0) = c(\cdot)$. Here, the set $G = K$ of problem (P_G) is supposed to be a cone, which was not necessarily the case in chapter 2.

We assume that, for some given point $x_0 \in \Omega$, the following smoothness properties of f and c hold:

- for all p in \mathcal{P} , $f(\cdot, p)$ and $c(\cdot, p)$ are differentiable on Ω ,
- $f(\cdot, p_0)$ and $c(\cdot, p_0)$ are twice continuously differentiable on Ω ,
- f , c , f'_x , and c'_x are continuous on $\Omega \times \mathcal{P}$.

We have denoted by f'_x and c'_x the derivatives of f and c , respectively, with respect to x . Similarly, we denote the second derivatives with respect to x of f and c by f''_{xx} and c''_{xx} , respectively.

From theorem 2.6 and proposition 2.21, it is known that at a local minimizer $x \in \Omega$ of (P_K^p) at which the following constraint qualification (CQ-R) holds

$$0 \in \text{int}(c(x, p) + c'_x(x, p)\mathbb{E} - K), \quad (3.1)$$

one can associate a multiplier $\lambda \in \mathbb{F}$ such that

$$\begin{cases} \nabla_x f(x, p) + c'_x(x, p)^* \lambda = 0 \\ K^- \ni \lambda \perp c(x, p) \in K. \end{cases} \quad (3.2)$$

Let us introduce the *multiplier multifunction* $\Lambda : \Omega \times \mathcal{P} \multimap \mathbb{F}$, which is defined at $(x, p) \in \Omega \times \mathcal{P}$ by

$$\Lambda(x, p) := \{\lambda \in \mathbb{F} : (x, p, \lambda) \text{ satisfies (3.2)}\},$$

as well as the *stationary point multifunction* $\Sigma : \mathcal{P} \multimap \Omega$, which is defined at $p \in \mathcal{P}$ by

$$\Sigma(p) := \{x \in \Omega : \Lambda(x, p) \neq \emptyset\}.$$

Proposition 3.1 (stability of (P_K) with a polyhedral K) *In the framework presented above, assume that*

- (i) \mathcal{P} is a subset of a vector space,
- (ii) K is a convex polyhedral cone,
- (iii) f'_x , c and c'_x are Lipschitz continuous near some $(x_0, p_0) \in \Omega \times \mathcal{P}$,
- (iv) (3.1) holds for $(x, p) = (x_0, p_0)$, and
- (v) there is a $\lambda_0 \in \Lambda(x_0, p_0)$ such that the strong SC2 holds for $(P_K^{p_0})$.

Then, there exists a constant $L \geq 0$, such that, for all p near p_0 , one has

- 1) $\Sigma(p) \neq \emptyset$,
- 2) for all $x \in \Sigma(p)$ near x_0 and all $\lambda \in \Lambda(x, p)$, one has

$$\text{dist}((x, \lambda), \{x_0\} \times \Lambda(x_0, p_0)) \leq L \|p - p_0\|.$$

Notes

Proposition 3.1 rephrases corollary 4.3 in [121; 1982].

4 A Few Methods for Nonsmooth Systems

This chapter presents and analyzes algorithms to solve various *nonsmooth* “systems” by “pseudo-linearization” techniques. The term *systems* is vague, but may be viewed as “mathematical models” in the discussion that follows; the types of systems considered in this chapter will be clarified in section 4.1.1; they include variational problems, variational inequality problems, complementarity problems, first order optimality conditions of an optimization problem, nonsmooth equations, to mention a few of the most important ones. The important qualifier *nonsmooth* means that these systems are defined by functions that are not differentiable in the classical sense of Fréchet, or are defined by multifunctions. The term *pseudo-linearization* refers to the fact that this lack of differentiability makes a true linearization impossible, but that some kind of surrogate is nevertheless available.

It may be instructive to gather the various approaches presented in this chapter into two classes. Both classes are constituted of algorithms that have a fast speed of convergence in a neighborhood of a *regular* solution (a concept to be defined, which depends on the considered system and algorithm), which is due to the pseudo-linearization of the systems they solve.

- We gather into the first class, the methods that try to have a rather precise description of the system to solve around the current iterate. The amount of information collected there gives to the iteration a nonlinear nature. This results in algorithms that have a rather complex iteration, which may require a rather significant computing effort to be processed (usually more than a single linear system to solve, certainly). An advantage of these algorithms is that they are often rather easy to *globalize*, meaning that one can design convergent versions of these algorithms even though the starting iterate is far from a solution.

The first example of methods of this class is the Josephy-Newton algorithm for solving functional inclusions. It is analyzed in section 4.1. Its rather precise description at the current iterate of the system to solve is obtained by keeping unaltered the part of the system that is not easy to linearize, the one that involves a multifunction.

The SQP algorithm of section 5.1 is also a member of the first class of approaches, since it can be viewed as an application of the Josephy-Newton algorithm to the first order optimality system of the optimization problem (P_{EI}) , which has equality and inequality constraints. The SQP algorithm is one of the most frequently used methods to solve (P_{EI}) . It was mainly developed in the mid-1970s.

The B-Newton algorithm of section 4.2, which solves a nonsmooth system of equations, has also a complex iteration. This one is due to the nonlinear nature of the

B-differential that is used to compute the new iterate. When applied to some reformulation of the first optimality conditions of a nonlinear optimization problem, however, it results in a simpler iteration than the SQP algorithm.

- We gather into the second class of methods, those that have a very economical iteration, in the sense that a single linear system needs to be solved at each iteration.

The semismooth algorithm of section 4.4, which solves a nonsmooth system of equations, belongs to this second class. The linear system to solve at the current iteration comes from a choice of local first order approximation of the nonsmooth system. Despite this poor description, a fast local convergence is also possible, which is a surprising and remarkable fact. On paper, the description of the algorithm is simple, but in practice the choice of the linear system is not always easy to determine. It may also be inappropriate, when the globalization of the method is an issue.

Interior point methods in linear, conic, or nonlinear optimization also belong to this class of methods, but we shall see them in another chapter, chapter ??.

4.1 Josephy-Newton Algorithm for Functional Inclusions

The Josephy-Newton (JN) algorithm has been designed to solve functional inclusions. This type of systems is described in section 4.1.1; they include variational problems (section 4.1.1), variational inequality problems (section 4.1.1), complementarity problems (section 4.1.1), and various systems of optimality conditions (the Peano-Kantorovich condition of the general problem (P_X) and the first order necessary optimality conditions of the problems (P_G) or (P_{EI}) ; see section 4.1.1). The algorithm is described in section 4.1.2, its asymptotic behavior is analyzed in section 4.1.4, and conditions ensuring its local convergence are provided in section 4.1.5.

4.1.1 A Gallery of Problems

Functional Inclusion Problem

Let \mathbb{E} and \mathbb{F} be two vector spaces having the same finite dimension. In this section, we are interested in solving a *functional inclusion problem*¹, which, by definition, is a problem that reads in the following manner:

$$(P_{FI}) \quad F(x) + N(x) \ni 0. \quad (4.1)$$

In this problem model, $F : \mathbb{E} \rightarrow \mathbb{F}$ is a function that is supposed to be differentiable and $N : \mathbb{E} \multimap \mathbb{F}$ is a multifunction “sufficiently simple”. This system means that one has to determine a point x in \mathbb{E} such that the set $F(x) + N(x)$ contains the zero element of \mathbb{F} ; one can also say that x is sought such that the vector $-F(x)$ is in the set $N(x)$. If N is the zero multifunction, one recovers a nonlinear system to solve

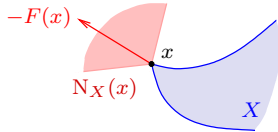
¹ This is the author’s own terminology. One more frequently finds the vaguer terms *generalized equation* [120, 46] or *variational inclusion* [2].

$F(x) = 0$. One could also incorporate F in N , hence removing the presence of the function value $F(x)$ in (P_{FI}) , without loss of generality, but the linearization method presented in section 4.1.2 takes advantage of this smooth function; furthermore, N could not be “simple” with this incorporation.

As we shall see, the system (P_{FI}) models many problems.

Variational Problem

A *variational problem*² is a functional inclusion problem of the form (P_{FI}) , in which $\mathbb{F} = \mathbb{E}$ and the multifunction $N : \mathbb{E} \multimap \mathbb{E}$ is the *normal cone map* N_X to a nonempty closed set $X \subseteq \mathbb{E}$ (the notation N in (P_{FI}) comes from this problem); in other words, $N_X(x)$ is the normal cone (1.50) to X at x . The problem reads and can be viewed as below:

$$(P_V) \quad F(x) + N_X(x) \ni 0. \tag{4.2}$$


With the convention that $N_x X = \emptyset$ when $x \notin X$, one looks for a point $x \in X$ such that $-F(x)$ is in the normal cone to X at x .

The first order necessary condition of optimality of Peano-Kantorovich (1.51) is a variational problem of the form (P_V) , in which F is the gradient of a function $f : \mathbb{E} \rightarrow \mathbb{R}$. However, unless the feasible set X is simple, its normal cone is often too complex for allowing the problem to be solved by the Josephy-Newton algorithm introduced in section 4.1.2 below.

Variational Inequality Problem

A *variational inequality problem* is a variational problem of the form (P_V) , in which the set $X \equiv C$ is nonempty, closed, and *convex*. Then, according to (1.23), problem (P_V) , which reads $-F(x) \in N_C(x)$, can also be written

$$(P_{VI}) \quad \begin{cases} x \in C \\ \langle F(x), y - x \rangle \geq 0, \quad \forall y \in C. \end{cases} \tag{4.3}$$


The presence of the infinite number of inequalities induces the problem name. Now, it is usually better not to see this problem as one with that infinite number of inequalities, but to see it in terms of multifunction like in (P_V) .

The system (4.3) can be written like a nonsmooth equation in x (see exercise 4.1.2)

$$P_C(x - F(x)) - x = 0. \tag{4.4}$$

This reformulation of the variational inequality problem is not necessarily advantageous, since the use of the projector P_C on C in (4.4) often yields a system that lacks appropriate differentiability properties (see remark 1.12), at a point that the usual

² One also encounters the denomination *variational condition* [126, 124].

techniques for solving equations cannot be used. This point of view may change for particular convex sets C .

An *equilibrium problem* is a significant extension of a variational inequality problem. Being given a set X of a topological space \mathbb{E} and a function $f : X \times X \rightarrow \mathbb{R}$, one seeks a point x such that

$$\begin{cases} x \in X \\ f(x, y) \geq 0, \quad \forall y \in X. \end{cases} \quad (4.5)$$

We will not talk about this problem in these notes.

Proposition 4.1 (existence of solution for (P_{VI})) Consider the variation inequality problem (P_{VI}) in (4.3) with a nonempty closed convex set C , on which F is continuous. Then, the set of solutions to (P_{VI}) is closed. This solution set is also nonempty if C is bounded.

Proof. □

Complementarity Problem

A *complementarity problem* is a functional inclusion problem of the form (P_{FI}) , in which \mathbb{F} is a Euclidean space and the multifunction N is defined at $x \in \mathbb{E}$ by $N(x) = N_{K^+}(G(x))$, where $G : \mathbb{E} \rightarrow \mathbb{F}$, K is a nonempty closed *convex cone* of \mathbb{F} , K^+ is the positive *dual cone* of K in \mathbb{F} , and N_{K^+} is the application “normal cone to K^+ ”. Then, by proceeding in the same way as for obtaining (P_{VI}) , the function inclusion becomes

$$G(x) \in K^+ \quad \text{and} \quad \left(\forall y \in K^+ : \langle F(x), y - G(x) \rangle \geq 0 \right).$$

Next, taking $y = 2G(x)$ and $y = G(x)/2$ (or zero, since K^+ is closed) as test elements in K^+ above, we see that the problem can be written

$$(P_C) \quad K^+ \ni G(x) \perp F(x) \in K. \quad (4.6)$$


This expression means that three conditions must be satisfied at the sought x : $F(x) \in K$, $G(x) \in K^+$, and $\langle F(x), G(x) \rangle_{\mathbb{F}} = 0$.

Constrained Optimization

Let \mathbb{E} and \mathbb{F} be two Euclidean vector spaces. Consider the problem of the form (P_G) :

$$\begin{cases} \min f(x) \\ c(x) \in K, \end{cases} \quad (4.7)$$

in which $f : \mathbb{E} \rightarrow \mathbb{R}$ and $c : \mathbb{E} \rightarrow \mathbb{F}$ are differentiable maps and K is a nonempty closed convex cone of \mathbb{F} . Its optimality conditions at a local solution $x \in \mathbb{E}$ is the following system in $(x, \lambda) \in \mathbb{E} \times \mathbb{F}$ (see theorem 2.6, in the case when $G \equiv K$ is a closed convex cone)

$$\nabla f(x) + c'(x)^* \lambda = 0 \quad \text{and} \quad K^- \ni \lambda \perp c(x) \in K. \quad (4.8)$$

This optimality system can also be written as a complementarity problem of the form (P_{CP}) , whose objects are topped by a tilde: the spaces are $\tilde{\mathbb{E}} = \tilde{\mathbb{F}} := \mathbb{E} \times \mathbb{F}$, the function $\tilde{G} := \mathbb{I}_{\tilde{\mathbb{E}}}$, while the function $\tilde{F} : \tilde{\mathbb{E}} \rightarrow \tilde{\mathbb{E}}$ and the set $\tilde{K} \subseteq \tilde{\mathbb{E}}$ are defined as follows

$$\tilde{F}(\tilde{x}) = \begin{pmatrix} \nabla f(x) + c'(x)^* \lambda \\ -c(x) \end{pmatrix} \quad \text{and} \quad \tilde{K} = \{0_{\mathbb{E}}\} \times (-K), \quad (4.9a)$$

where we have set $\tilde{x} := (x, \lambda)$. Since $\tilde{K}^+ = \mathbb{E} \times K^-$, the system (4.8) reads

$$\tilde{K}^+ \ni \tilde{x} \perp \tilde{F}(\tilde{x}) \in \tilde{K}, \quad (4.9b)$$

which is of the form of (P_{CP}) in (4.6).

Since the optimization problem (P_{EI}) , with equality and inequality constraints, can be written like (4.7) with $\mathbb{F} = \mathbb{R}^{m_E} \times \mathbb{R}^{m_I}$, $c = (c_E, c_I) : \mathbb{E} \rightarrow \mathbb{R}^{m_E} \times \mathbb{R}^{m_I}$, and $K = \{0_{\mathbb{R}^{m_E}}\} \times \mathbb{R}_+^{m_I}$, its optimality conditions, the KKT system (1.58), can also be written as a complementarity problem (4.9b). From (4.9a), it suffices to take $\tilde{x} := (x, \lambda)$ and

$$\tilde{F}(\tilde{x}) := \begin{pmatrix} \nabla f(x) + c'(x)^* \lambda \\ -c(x) \end{pmatrix} \quad \text{and} \quad \tilde{K} = \{0_{\mathbb{E}}\} \times (\{0_{\mathbb{R}^{m_E}}\} \times \mathbb{R}_+^{m_I}). \quad (4.10)$$

Equality and Inequality Systems

When N is the constant multifunction $x \in \mathbb{E} \mapsto \{0_{\mathbb{R}^{m_E}}\} \times \mathbb{R}_+^{m_I} \subseteq \mathbb{R}^m \equiv \mathbb{F}$, where E and I form a partition of $[1 : m]$, the functional inclusion problem (P_{FI}) amounts to find a point $x \in \mathbb{E}$ satisfying a system of equalities and inequalities:

$$F_E(x) = 0 \quad \text{and} \quad F_I(x) \leq 0.$$

Nevertheless, as we shall see, the algorithmic approach of section 4.1 will not be immediately workable to find a solution to such a system, because its solutions are usually not isolated.

4.1.2 The Josephy-Newton Algorithm

The *Josephy-Newton* (JN) *algorithm* is an iterative method to solve the functional inclusion problem (P_{FI}) in (4.1). It generates a sequence $\{x_k\} \subseteq \mathbb{E}$ as follows. Knowing $x_k \in \mathbb{E}$, the next iterate $x_{k+1} \in \mathbb{E}$ is computed as a solution in x (if this is possible!) to the functional inclusion (P_{FI}) , “linearized” in x , namely

$$F(x_k) + F'(x_k) \cdot (x - x_k) + N(x) \ni 0. \quad (4.11)$$

The linearization is only partial: whilst F is linearized at x_k like in Newton’s algorithm (1.68), the multifunction N is left unchanged. Such a linearization of (P_{FI}) is named a

JN linearization below. The system (4.11) to solve in x_{k+1} at each iteration is therefore often “nonlinear”, meaning that it is not enough to solve a single linear system to get one of its solutions. As a result, the algorithm is generally more expensive than the standard Newton algorithm (1.68) for the system of equations (1.67). One can also say that the algorithm captures from the system, much more than the linear approximation of its functions, since N is used without approximation. For future references, let us write explicitly the functional inclusion verified by $x_{k+1} \in \mathbb{E}$:

$$F(x_k) + F'(x_k) \cdot (x_{k+1} - x_k) + N(x_{k+1}) \ni 0. \tag{4.12}$$

Now, it may arise that no algorithm is known to solve (4.11) efficiently; in which case, it is necessary to turn towards other solution approaches.

One can now understand the assumptions required on F and N after (4.1). The function F must be differentiable so that its derivative operator $F'(x_k)$ can be used in the definition (4.12) of the algorithm. The multifunction N needs not have differentiability properties, but must be simple enough to make the computation of a solution to (4.12) not too expensive.

4.1.3 Regularity

The standard quadratic convergence result of Newton’s method for finding a zero of a nonlinear system of equations $F(x) = 0$, the one of theorem 1.48, requires the nonsingularity of $F'(x_*)$. We would like to extend this assumption to the more complex functional inclusion system $(P_{\mathbb{F}\mathbb{I}})$ in (4.1), which is not straightforward with the presence of the multifunction N , in order to ensure quadratic convergence of the Josephy-Newton algorithm. Getting such an extension is often a matter of rephrasing.

Consider the system $F(x) = 0$, for which we rephrase the property of the injectivity of $F'(x_*)$, which is equivalent to the nonsingularity of $F'(x_*)$ when $\dim \mathbb{E} = \dim \mathbb{F}$. Observe that, when $F(x_*) = 0$, the definition of the assumed differentiability of F implies that

$$F(x) = F'(x_*)(x - x_*) + o(\|x - x_*\|).$$

Now, if $F'(x_*)$ is injective, this development implies that, for some constant $\sigma_1 > 0$ and $\sigma_2 > 0$, independent of x ,

$$\|x - x_*\| \leq \sigma_1 \quad \implies \quad \|x - x_*\| \leq \sigma_2 \|F(x)\|.$$

This property can also be written as follows: there are constants $\sigma_1 > 0$ and $\sigma_2 > 0$ such that

$$\left. \begin{array}{l} (x, p) \in \mathbb{E} \times \mathbb{F} \\ F(x) = p \\ \|x - x_*\| \leq \sigma_1 \end{array} \right\} \implies \|x - x_*\| \leq \sigma_2 \|p\|.$$

It is this formulation of the injectivity of $F'(x_*)$ that we chose to keep as a desired property of a root x_* of the functional inclusion system, because it makes no use of the differential of F at the considered zero x_* and can therefore be applied to $F + N$. This concept is called semi-stability; it was introduced in [19].

Definition 4.2 (semi-stability) A solution x_* to the functional inclusion (P_{FI}) in (4.1) is said to be *semi-stable* if there are constants $\sigma_1 > 0$ and $\sigma_2 > 0$ such that

$$\left. \begin{array}{l} (x, p) \in \mathbb{E} \times \mathbb{F} \\ F(x) + N(x) \ni p \\ \|x - x_*\| \leq \sigma_1 \end{array} \right\} \implies \|x - x_*\| \leq \sigma_2 \|p\|. \quad \square$$

In plain words, one can also say that x_* is semi-stable if any solution x of the slightly perturbed system $F(x) + N(x) \ni p$ that is close to x_* , depends on the perturbation p in a Lipschitz manner.

Remarks 4.3

- 1) The semi-stability is a concept that applies on the “nonlinear” functional inclusion $F(x) + N(x) \ni 0$, not on its “linearization” $F(x_*) + F'(x_*)(x - x_*) + N(x') \ni 0$. By proposition 4.6 below, we shall see, however, that the semi-stability implies a property on the linearized functional inclusion and that this property is actually equivalent to the semi-stability if N is the normal cone to a convex polyhedron (proposition 4.7).
- 2) The semi-stability does not claim anything on the existence of solution to the functional inclusion $F(x) + N(x) \ni p$ for the considered (x, p) . It only requires a Lipschitz property with respect to the perturbation of the solutions to the perturbed system $F(x) + N(x) \ni p$, if these solutions exist and are close to x_* .
- 3) The straightforward following observation will have important consequences below.

Proposition 4.4 (isolation of a semi-stable solution) *A semi-stable solution to (P_{FI}) is isolated.*

Proof. It is certainly sufficient to show that, if x_* is a semi-stable solution to (P_{FI}) , the functional inclusion has no other solution than x_* in $\bar{B}(x_*, \sigma_1)$, where $\sigma_1 > 0$ is the constant appearing in the definition of the semi-stability. Let $x'_* \in \bar{B}(x_*, \sigma_1)$ be a solution to (P_{FI}) . Hence $\|x'_* - x_*\| \leq \sigma_1$ and $F(x'_*) + N(x'_*) \ni 0$, so that $\|x'_* - x_*\| \leq \sigma_2 \|0\|$, implying that $x'_* = x_*$. □

- 4) Semi-stability can also be viewed as a *local error bound* for the isolated solution set $\{x_*\}$, since it also reads: there exist positive constants σ_1 and σ_2 such that

$$\|x - x_*\| \leq \sigma_1 \implies \|x - x_*\| \leq \sigma_2 \text{dist}(0, F(x) + N(x)). \quad (4.13)$$

Indeed, taking the infimum in $p \in F(x) + N(x)$ in the definition 4.2 of the semi-stability yields the implication above. Conversely, this implication yields the one in definition 4.2 since $\text{dist}(0, F(x) + N(x)) \leq \|p\|$ when $\text{dist}(0, F(x) + N(x)) \ni p$.

- 5) As expected by its construction, the notion of semi-stability reduces to the injectivity of $F'(x_*)$ in the absence of the multifunction N (hence to its nonsingularity when $\dim \mathbb{E} = \dim \mathbb{F}$).

Proposition 4.5 (semi-stability and nonsingularity) *Assume that $N \equiv 0$ in (P_{FI}) and that F is differentiable at a point x_* such that $F(x_*) = 0$. Then x_* is a semi-stable solution to (P_{FI}) if and only if $F'(x_*)$ is injective.*

Proof. [\Rightarrow] Let d such that $F'(x_*)d = 0$ and set $x_t := x_* + td$ with $t > 0$. Then $F(x_t) = o(t)$ by the differentiability of F at x_* and $F(x_*) = 0$. The semi-stability of x_* then implies that, as soon as $t > 0$ is small enough, $t\|d\| = \|x_t - x_*\| \leq \sigma_2 o(t)$. Hence, one must have $d = 0$.

[\Leftarrow] This is a consequence of the reasoning that led to the definition of the semi-stability (see the discussion before definition 4.2). By the differentiability of F at x_* and $F(x_*) = 0$, one has

$$F(x) = F'(x_*)(x - x_*) + o(\|x - x_*\|). \tag{4.14}$$

By the injectivity of $F'(x_*)$, there is a constant $C > 0$ such that

$$\|F'(x_*)h\| \geq C\|h\|, \quad \forall h \in \mathbb{E}. \tag{4.15}$$

Furthermore, by (4.14), there exists $\sigma_1 > 0$ such that

$$\|x - x_*\| \leq \sigma_1 \implies \|F(x) - F'(x_*)(x - x_*)\| \geq \frac{C}{2}\|x - x_*\|. \tag{4.16}$$

Combining (4.14), (4.15), and (4.16), we obtain

$$\|x - x_*\| \leq \sigma_1 \implies \|F(x)\| \geq \frac{C}{2}\|x - x_*\|.$$

Setting $\sigma_2 = 2/C$, we see that if $(x, p) \in \mathbb{E} \times \mathbb{F}$ verifies $\|x - x_*\| \leq \sigma_1$ and $F(x) = p$, the following holds $\|x - x_*\| \leq \sigma_2\|p\|$. \square

We have observed that a semi-stable solution to the functional inclusion (P_{FI}) is an isolated solution (proposition 4.4). The next result tells us that it is also an isolated solution to the functional inclusion, derived from (P_{FI}) by linearizing F at x_* :

$$F(x_*) + F'(x_*)(x - x_*) + N(x) \ni 0. \tag{4.17}$$

We shall see in proposition 4.7 that the reciprocal holds when N is the multifunction “normal cone to a convex polyhedron”.

Proposition 4.6 (isolated solution to the linearized functional inclusion) *Let x_* be a semi-stable solution to the functional inclusion (P_{FI}) , in which F is differentiable at x_* . Then, x_* is an isolated solution to (4.17).*

Proof. Let $\sigma_1 > 0$ and $\sigma_2 > 0$ be the constants given by the semi-stability of x_* . Suppose that there is a solution x to (4.17) that is arbitrary close to x_* . Then,

$$F(x) + N_C(x) \ni F(x) - F(x_*) - F'(x_*)(x - x_*) \quad \text{and} \quad \|x - x_*\| \leq \sigma_1.$$

By the semi-stability and the differentiability of F at x_* , the following holds

$$\|x - x_*\| \leq \sigma_2 \|F(x) - F(x_*) - F'(x_*)(x - x_*)\| = o(\|x - x_*\|).$$

This implies that $x = x_*$. Hence a solution close to x_* cannot be different from x_* . \square

We understand that the semi-stability property of a solution x_* to the functional inclusion problem (P_{FI}) is rather restrictive, since it requires in particular that

- x_* be an isolated solution to the functional inclusion (P_{FI}) (proposition 4.4),
- x_* be an isolated solution to (4.17), the functional inclusion linearized at x_* (proposition 4.6).

One of the interests of semi-stability is to identify the situations where the JN algorithm described and analyzed in section 4.1 has fast convergence (proposition 4.9 and theorem 4.12) and it is known that linearization algorithms, like Newton's method (section 1.5.2) or the JN algorithm (section 4.1.2) or the SQP algorithm (section 5.1), can behave badly when the aimed solution is not isolated.

We conclude this section with characterizations of the semi-stability, when N is the map “normal cone N_C to a nonempty convex polyhedron C ”, that highlight the role of the linearized functional inclusion. The condition (ii) will be used for characterizing the semi-stability of a stationary point of (P_{EI}) (proposition ??) and the condition (iii) for characterizing the semi-stability of a local solution to (P_{EI}) (proposition 5.4). Be mindful that the normal cone is evaluated at the point x_* in (iii), while this evaluation is done at x in (ii) and (iv).

Proposition 4.7 (characterization of the semi-stability of a polyhedral

VI) *Suppose that $\mathbb{E} = \mathbb{F}$ and that $x_* \in \mathbb{E}$ is a solution to the functional inclusion (P_{FI}) , in which F is differentiable at x_* and N is the map normal cone N_C to a nonempty closed convex set C of \mathbb{E} . Then, the implications (i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (iv) hold for the following claims:*

- (i) x_* is semi-stable,
- (ii) x_* is an isolated solution to the linearized functional inclusion

$$F(x_*) + F'(x_*)(x - x_*) + N_C(x) \ni 0, \quad (4.18)$$

- (iii) any $x \in C \setminus \{x_*\}$ verifying

$$\langle F(x_*), x - x_* \rangle = 0 \quad (4.19a)$$

$$F(x_*) + F'(x_*)(x - x_*) + N_C(x_*) \ni 0, \quad (4.19b)$$

is such that $\langle F'(x_)(x - x_*), x - x_* \rangle > 0$,*

- (iv) the system in x below has no other solution than x_* :

$$N_C(x) \subseteq N_C(x_*) \quad (4.20a)$$

$$\langle F(x_*), x - x_* \rangle = 0 \quad (4.20b)$$

$$\mathbb{R}_+ F(x_*) + F'(x_*)(x - x_*) + N_C(x) \ni 0. \quad (4.20c)$$

If C is a nonempty convex polyhedron, then the four properties (i)-(iv) are equivalent.

Proof. [(i) \Rightarrow (ii)] This is proposition 4.6 for the particular case when $N = N_C$.

[(ii) \Rightarrow (iii)] We show the contrapositive, assuming that there exists an $x_1 \in C \setminus \{x_*\}$ such that (iii) does not hold or

$$\langle F(x_*), x_1 - x_* \rangle = 0, \quad (4.21a)$$

$$F(x_*) + F'(x_*)(x_1 - x_*) + N_C(x_*) \ni 0, \quad (4.21b)$$

$$\langle F'(x_*)(x_1 - x_*), x_1 - x_* \rangle \leq 0 \quad (4.21c)$$

and show that $x_t := (1-t)x_* + tx_1$ is a solution to (4.18) for all $t \in [0, 1]$, which contradicts (ii).

Observe first that one has an equality in (4.21c):

$$\langle F'(x_*)(x_1 - x_*), x_1 - x_* \rangle = 0. \quad (4.22)$$

This is because by (4.21b) and $x_1 \in C$, one has $\langle F(x_*) + F'(x_*)(x_1 - x_*), x_1 - x_* \rangle \geq 0$. Next, using (4.21a) in this inequality yields $\langle F'(x_*)(x_1 - x_*), x_1 - x_* \rangle \geq 0$, which is the reverse inequality of (4.21c), hence (4.22).

To show that x_t is a solution to (4.18), we have to prove that $x_t \in C$, which is clear by the convexity of C , and that, for all $y \in C$, the following value is nonnegative:

$$\begin{aligned} & \langle F(x_*) + F'(x_*)(x_t - x_*), y - x_t \rangle \\ &= \langle F(x_*) + tF'(x_*)(x_1 - x_*), y - x_* - t(x_1 - x_*) \rangle \\ &= \langle F(x_*) + tF'(x_*)(x_1 - x_*), y - x_* \rangle \quad [(4.21a) \text{ and } (4.22)] \\ &\geq (1-t)\langle F(x_*), y - x_* \rangle \quad [(4.21b)] \\ &\geq 0 \quad [1-t \geq 0 \text{ and } F(x_*) + N_C(x_*) \ni 0]. \end{aligned}$$

[(iii) \Rightarrow (iv)] We show the contrapositive: assuming that there exists an $x \neq x_*$ satisfying (4.20), we find some $t > 0$ such that $x_t := x_* + t(x - x_*)$ satisfies

$$\langle F(x_*), x_t - x_* \rangle = 0, \quad (4.23a)$$

$$F(x_*) + F'(x_*)(x_t - x_*) + N_C(x_*) \ni 0, \quad (4.23b)$$

$$\langle F'(x_*)(x_t - x_*), x_t - x_* \rangle \leq 0, \quad (4.23c)$$

which contradicts (4.19).

Observe already that (4.23a) follows at once from (4.20b). To get (4.23c), we use (4.20c), which reads for some $\alpha \geq 0$:

$$\langle \alpha F(x_*) + F'(x_*)(x - x_*), y - x \rangle \geq 0, \quad \forall y \in C. \quad (4.24)$$

Now (4.23c) follows by taking $y = x_* \in C$ in this inequality, multiplying the left-hand side by t^2 (hence $x - x_*$ becomes $x_t - x_*$) and using (4.23a). It remains to show (4.23b), which also reads

$$\langle F(x_*) + F'(x_*)(x_t - x_*), y - x_* \rangle \geq 0, \quad \forall y \in C. \quad (4.25)$$

This is almost the form (4.24) of (4.20c). We examine two cases.

- If $\alpha > 1$, we divide the left-hand side of (4.24) by α and get (4.25) with $t = 1/\alpha \in (0, 1]$.
- If $\alpha \in [0, 1]$, instead of (4.24), we use (4.20a) and (4.20c) to get for some $\alpha \geq 0$:

$$\langle \alpha F(x_*) + F'(x_*)(x - x_*), y - x_* \rangle \geq 0, \quad \forall y \in C. \quad (4.26)$$

Now, $F(x_*) + N_C(x_*) \ni 0$ (x_* is a solution to (P_{VI})) implies that

$$\langle F(x_*), y - x_* \rangle \geq 0, \quad \forall y \in C.$$

Multiplying the left-hand side of these inequalities by $1 - \alpha \geq 0$ and combining with (4.26), yield (4.25) with $t = 1$.

[(iv) \Rightarrow (i)] Suppose now that C is a convex polyhedron. We show the contrapositive, assuming that x_* is not semi-stable and show that the system (4.20) has a solution $x \neq x_*$.

Letting the constants $\sigma_1 \rightarrow 0$ and $\sigma_2 \rightarrow \infty$ in the definition 4.2 of the semi-stability, we see that one can find sequences $\{x_k\} \subseteq \mathbb{E}$ and $\{p_k\} \subseteq \mathbb{F}$ such that

$$F(x_k) + N_C(x_k) \ni p_k, \quad (4.27a)$$

$$x_k \rightarrow x_* \quad \text{with} \quad x_k \neq x_*, \quad (4.27b)$$

$$\|p_k\|/\|x_k - x_*\| \rightarrow 0. \quad (4.27c)$$

Extracting a subsequence if necessary, one can suppose that, with $t_k := \|x_k - x_*\|$, we have $(x_k - x_*)/t_k \rightarrow d$. Since $x_k \in C$ by (4.27a), it clearly follows that $d \in T_C(x_*) \setminus \{0\}$.

Let us now take the limit in (4.27a), after the expansion of $F(x_k)$ around x_* and division by $t_k > 0$:

$$\frac{1}{t_k} F(x_*) + F'(x_*) \frac{x_k - x_*}{t_k} + \frac{o(\|x_k - x_*\|)}{t_k} + N_C(x_k) \ni \frac{p_k}{t_k}.$$

The first term is annoying since $t_k \rightarrow 0$, but it belongs to $\mathbb{R}_+ F(x_*)$ and we can replace it by that set, while keeping the inclusion. The second term tends to $F'(x_*)d$ and the third one to zero. By extracting a subsequence if necessary, one can fix the normal cones in the fourth term to a unique one (exercise 1.2.1), which is denoted by $N_C(x_k) \equiv N_C(x_0)$ below. Finally, the right-hand side tends to zero by (4.27c). Since $\mathbb{R}_+ F(x_*) + N_C(x_0)$ is closed (the sum of two convex polyhedrons is a polyhedron [a property recalled in point 2 of proposition 1.1], hence closed), we obtain at the limit:

$$\mathbb{R}_+ F(x_*) + F'(x_*)d + N_C(x_0) \ni 0.$$

This inclusion resembles (4.20c), which we now try to establish for the point

$$x = x_* + \varepsilon d,$$

where $\varepsilon > 0$ is taken sufficiently small in order to have $x \in C \setminus \{x_*\}$; this is possible since $d \neq 0$ and $d \in T_C(x_*) = T_C^f(x_*)$ by (1.26a). With that x , the preceding inclusion becomes

$$\mathbb{R}_+ F(x_*) + F'(x_*)(x - x_*) + N_C(x_0) \ni 0. \quad (4.28)$$

For $y \in C$, let us denote by $I(y)$ the set of the indices of the active inequality affine constraints defining the polyhedron C that are active at y . To get (4.20) and therefore conclude, it is sufficient to show that

$$I(x_0) \subseteq I(x) \subseteq I(x_*), \quad (4.29a)$$

$$\langle F(x_*), x - x_* \rangle = 0. \quad (4.29b)$$

Indded, by (1.27b), (4.29a) implies that $N_C(x_0) \subseteq N_C(x) \subseteq N_C(x_*)$ and therefore (4.20a); (4.29b) is (4.20b); and (4.28) and $N_C(x_0) \subseteq N_C(x)$ implies (4.20c).

Consider first (4.29a). Observe that the convergence $x_k \rightarrow x_*$ and the fact that $I(x_k) \equiv I(x_0)$ (for the subsequence $\{x_k\}$ having the same $I(x_k)$, selected above) imply that

$$I(x_k) = I(x_0) \subseteq I(x_*).$$

Therefore, if $i \in I(x_k)$, it follows that $i \in I(x_*)$ and that $i \in I((1 - \varepsilon/t_k)x_* + (\varepsilon/t_k)x_k)$. Since $([1 - \varepsilon/t_k]x_* + [\varepsilon/t_k]x_k) = x_* + \varepsilon(x_k - x_*)/t_k \rightarrow x_* + \varepsilon d = x$, we have the first inclusion in (4.29a). For the second inclusion in (4.29a), we take $\varepsilon > 0$ small enough, so that $x = x_* + \varepsilon d$ is sufficiently close to x_* to have $I(x) \subseteq I(x_*)$.

Consider now (4.29b). First, $\langle F(x_*), x - x_* \rangle \geq 0$, since $F(x_*) + N_C(x_*) \ni 0$ and $x \in C$. To get the reverse inequality, we use (4.27a) and $x_* \in C$ to get $\langle F(x_k) - p_k, x_* - x_k \rangle \geq 0$. Dividing by $t_k > 0$ and taking the limit in k yield $\langle F(x_*), d \rangle \leq 0$ since $(x_* - x_k)/t_k \rightarrow -d$, hence $\langle F(x_*), x - x_* \rangle \leq 0$ since $d = (x - x_*)/\varepsilon$. \square

Definition 4.8 (strong regularity) A solution x_* to the functional inclusion (P_{FI}) is said to be *strongly regular* if there is a constant $\varepsilon > 0$ such that, for all p near zero, the system

$$\begin{cases} F(x_*) + F'(x_*)(x - x_*) + N(x) \ni p \\ \|x - x_*\| \leq \varepsilon \end{cases}$$

has a unique solution $x(p)$ and $x(\cdot)$ is Lipschitz near zero. \square

This is a much stronger assumption than semi-stability in that it assumes the existence of a solution for small perturbations p and that this solution is unique.

4.1.4 Speed of Convergence

In this section, we consider a structurally slight but important generalization of the JN algorithm (4.12), in which the next iterate x_{k+1} following the current one x_k is computed by

$$F(x_k) + M_k(x_{k+1} - x_k) + N(x_{k+1}) \ni 0, \quad (4.30)$$

where $M_k \in \mathcal{L}(\mathbb{E}, \mathbb{F})$ may be the Jacobian $F'(x_k)$ or an approximation to it. Hence this algorithm includes the quasi-Newton versions of the JN algorithm.

Let us now highlight conditions ensuring superlinear and quadratic convergence of the sequences generated by the JN algorithm (4.30). In this analysis, we assume that a sequence $\{x_k\}$ is generated by the algorithm and that this sequence converges to a solution x_* to (P_{FI}) . These conditions depend on the smoothness of F , on the quality of the approximation of $F'(x_k)$ by M_k , and on the regularity of the solution x_* (its semi-stability, actually). The conditions on M_k used in proposition 4.9 below have to be compared with the Dennis and Moré condition for nonlinear systems (proposition 1.49): fast convergence is guaranteed, provided M_k is close to $F'(x_k)$ along the displacement direction $x_{k+1} - x_k$. Nothing is required on the multifunction N ; a posteriori, this can be understood by the fact that the algorithm takes all information from N , not only a kind of linearization like it does for F .

Proposition 4.9 (sufficient conditions for fast convergence) *Let x_* be a semi-stable solution to (P_{FI}) . Suppose that F is differentiable in a neighborhood of x_* and that F' is continuous at x_* . Let $\{x_k\}$ be a sequence satisfying the recurrence (4.30) and converging to x_* .*

- 1) *If $(M_k - F'(x_*))(x_{k+1} - x_k) = o(\|x_{k+1} - x_k\|)$, then the convergence of $\{x_k\}$ is superlinear.*
- 2) *If $(M_k - F'(x_*))(x_{k+1} - x_k) = O(\|x_{k+1} - x_k\|^2)$ and if F' is radially Lipschitz at x_* , then the convergence of $\{x_k\}$ is quadratic.*

Proof. Let σ_1 and σ_2 be the positive constants of the semi-stable solution x_* and let us simplify the writing by introducing $s_k := x_{k+1} - x_k$ and $\Delta_k := (M_k - F'(x_*))s_k$.

0) We want to apply the semi-stability to have an estimate of the updated error $x_{k+1} - x_*$, hence having an implication of the form

$$\left. \begin{array}{l} F(x_{k+1}) + N(x_{k+1}) \ni p_{k+1} \\ \|x_{k+1} - x_*\| \leq \sigma_1 \end{array} \right\} \implies \|x_{k+1} - x_*\| \leq \sigma_2 \|p_{k+1}\|. \quad (4.31)$$

The fact that $\|x_{k+1} - x_*\| \leq \sigma_1$ is guaranteed for large k , by the assumed convergence of $\{x_k\}$ to x_* . For the inclusion in the left-hand side of (4.31), it is natural to start with the iteration recurrence (4.30), which provides the aforementioned inclusion with

$$p_{k+1} := F(x_{k+1}) - [F(x_k) + M_k s_k] = F(x_{k+1}) - F(x_k) - F'(x_*)s_k - \Delta_k.$$

Therefore, by the implication (4.31), we have

$$\|x_{k+1} - x_*\| \leq \sigma_2 \|p_{k+1}\|. \quad (4.32)$$

The goal now is to get an estimate of $\|p_{k+1}\|$ in terms of $\|x_k - x_*\|$. By the mean value theorem (corollary 1.25), we have that

$$\|p_{k+1}\| \leq \left(\sup_{t \in [0,1]} \|F'(x_k + ts_k) - F'(x_*)\| \right) \|s_k\| - \Delta_k. \quad (4.33)$$

1) The continuity of F' at x_* and the assumption $\Delta_k = o(\|s_k\|)$ of the case allow us to deduce from (4.33) that $p_{k+1} = o(\|s_k\|)$. Therefore $x_{k+1} - x_* = o(\|s_k\|)$ by

(4.32). We deduce from this estimate that $x_{k+1} - x_* = o(\|x_k - x_*\|)$, which is the mark of the superlinear convergence of $\{x_k\}$.

2) If F' is L -Lipschitz continuous near x_* and $\Delta_k = O(\|s_k\|^2)$, we can estimate p_{k+1} from (4.33) as follows

$$\|p_{k+1}\| \leq L\|x_k - x_*\| \|s_k\| + O(\|s_k\|^2) = O(\|x_k - x_*\|^2),$$

since $s_k \sim (x_k - x_*)$, by the superlinear convergence of $\{x_k\}$ established in point 1 and lemma 1.47. Then, one deduces from (4.32) that $x_{k+1} - x_* = O(\|x_k - x_*\|^2)$, which is the mark of the quadratic convergence of $\{x_k\}$. \square

The next corollary, whose proof is straightforward, essentially deals with the case when $M_k = F'(x_k)$.

Corollary 4.10 (speed of convergence of the JN algorithm) *Suppose that F is C^1 in a neighborhood of a semi-stable solution x_* to (P_{FI}) and that the sequence $\{x_k\}$ satisfies the recurrence (4.11) and converges to x_* .*

- 1) *If $M_k \rightarrow F'(x_*)$, then the convergence of $\{x_k\}$ is superlinear.*
- 2) *If $M_k - F'(x_*) = O(\|x_k - x_*\|)$ and if F' is **radially Lipschitz** at x_* , then the convergence of $\{x_k\}$ is quadratic.*

Proof. 1) Point 1 follows immediately from point 1 of proposition 4.9.

2) If $M_k - F'(x_*) = O(\|x_k - x_*\|)$, then $(M_k - F'(x_*))(x_{k+1} - x_k) = O(\|x_k - x_*\| \|x_{k+1} - x_k\|)$, which implies the superlinear convergence of $\{x_k\}$ by point 1 of proposition 4.9. Next lemma 1.47 implies that $(x_k - x_*) \sim (x_{k+1} - x_k)$, from which we have $(M_k - F'(x_*))(x_{k+1} - x_k) = O(\|x_{k+1} - x_k\|^2)$. Now the quadratic convergence of $\{x_k\}$ follows from the **radial Lipschitz continuity** of F' at x_* and point 2 of proposition 4.9. \square

4.1.5 Local Convergence

The preceding section analyzed the speed of convergence of the sequence of iterates $\{x_k\}$ generated by the JN algorithm, assuming that such a sequence is generated and that this one converges to some semi-stable solution to the functional inclusion problem (P_{FI}) in (4.1). This section clarifies the last two aspects. On the one hand, the well-posedness of the algorithm is shown, which amounts to ensure that the linearized functional inclusion (4.11) has a solution. On the other hand, the local convergence of the algorithm is proved, meaning that the generated sequence converges to the considered neighboring solution x_* . For the two goals, the current iterate x_k is supposed to be close to a solution x_* , having an additional property.

Whilst the **semi-stability** of definition 4.2 has a local injectivity flavor, revealed by propositions 4.4 and 4.5, the **hemi-stability** defined below has a local surjectivity interpretation, in the sense that it requires the pseudo-linearized function inclusion to have a solution at linearization points close to x_* .

Definition 4.11 (hemi-stability) A solution x_* to (P_{F1}) is said to *hemi-stable* if for all $\alpha > 0$, there exists $\beta > 0$ such that, for all $x_0 \in \bar{B}(x_*, \beta)$, the following functional inclusion in x

$$F(x_0) + F'(x_0)(x - x_0) + N(x) \ni 0$$

has a solution in $\bar{B}(x_*, \alpha)$. □

In plain words, x_* is hemi-stable if one can find a solution to the linearized functional inclusion (4.11) that is as close as desired to x_* , just by taking the point of linearization x_k sufficiently close to x_* . It is not claimed that, when x_k is close to x_* , the linearized functional inclusion (4.11) has a unique solution; in particular, this one could have another solution that is not close to x_* . In the next theorem, we use the phrase “can generate” to express the fact that, when x_k is close to an hemi-stable solution, the JN algorithm takes a solution x_{k+1} to the linearized inclusion (4.11) that is also close to the solution x_* .

Theorem 4.12 (local convergence of the JN algorithm) *Let x_* be a semi-stable and hemi-stable solution to (P_{F1}) . Suppose that F is differentiable in a neighborhood of x_* and that F' is continuous at x_* . Consider the JN algorithm (4.12). Then, there exists an $\varepsilon > 0$, such that*

- 1) *if the first iterate x_1 is in the closed ball $\bar{B}(x_*, \varepsilon)$, then the JN algorithm can generate a sequence $\{x_k\}$ in $\bar{B}(x_*, \varepsilon)$,*
- 2) *any sequence $\{x_k\}$ generated in $\bar{B}(x_*, \varepsilon)$ by the JN algorithm converges superlinearly to x_* (and quadratically if F' is radially Lipschitz at x_*).*

Proof. 0) Let us first fix a few constants in the right order. Let $\sigma_1 > 0$ and $\sigma_2 > 0$ be the constants given by the semi-stability at x_* . Define

$$p(x, x') := F(x') - F(x) - F'(x)(x' - x).$$

The mean value theorem (corollary 1.25) allows us to write

$$\|p(x, x')\| \leq \left(\sup_{z \in (x, x')} \|F'(z) - F'(x)\| \right) \|x' - x\|.$$

By the assumed continuity of F' at x_* , the factor in parentheses can be made as small as desired by taking x and x' close enough to x_* . Therefore, one can find a constant $\alpha \in (0, \sigma_1]$ such that

$$x, x' \in \bar{B}(x_*, \alpha) \implies \|p(x, x')\| \leq \frac{1}{3\sigma_2} \|x' - x\|. \tag{4.34}$$

Let $\beta > 0$ be the constant associated with α by the hemi-stability of x_* and set $\varepsilon := \min(\alpha, \beta)$.

1) Suppose now that $x_k \in \bar{B}(x_*, \varepsilon)$. By the hemi-stability of x_* and $\varepsilon \leq \beta$, there exists a new iterate $x_{k+1} \in \bar{B}(x_*, \alpha)$ such that

$$F(x_k) + F'(x_k)(x_{k+1} - x_k) + N(x_{k+1}) \ni 0. \tag{4.35}$$

Since this inclusion is the **JN** algorithm recurrence formula (4.12), the algorithm *can* take this x_{k+1} as the iterate following x_k . Therefore, we have shown that the **JN** algorithm *can* generate a sequence in $\bar{B}(x_*, \varepsilon)$.

2) Assume now that the **JN** generates a sequence $\{x_k\}$ in $\bar{B}(x_*, \varepsilon)$. Hence the inclusion (4.35) holds. Let us now apply the *semi-stability* property of x_* . It follows from (4.35) that

$$F(x_{k+1}) + N(x_{k+1}) \ni F(x_{k+1}) - F(x_k) - F'(x_k)(x_{k+1} - x_k) = p(x_k, x_{k+1}).$$

Since $x_{k+1} \in \bar{B}(x_*, \varepsilon)$ and $\varepsilon \leq \alpha \leq \sigma_1$, the semi-stability at x_* implies that

$$\begin{aligned} \|x_{k+1} - x_*\| &\leq \sigma_2 \|p(x_k, x_{k+1})\| \\ &\leq \frac{1}{3} \|x_{k+1} - x_k\| \quad [(4.34) \text{ and } x_k, x_{k+1} \in \bar{B}(x_*, \alpha)] \\ &\leq \frac{1}{3} \|x_{k+1} - x_*\| + \frac{1}{3} \|x_k - x_*\|. \end{aligned}$$

Therefore $\|x_{k+1} - x_*\| \leq \frac{1}{2} \|x_k - x_*\|$. We have shown that $x_k \rightarrow x_*$.

The superlinear and quadratic convergence speed of convergence follows from corollary 4.10. \square

4.1.6 Globalization by Line-Search \blacktriangle

Notes

Historically, the variational inequality problem (P_{VI}) was introduced by Hartman and Stampacchia [64; 1966] for solving some nonlinear elliptic partial differential equation and was subsequently developed in many papers, including [94, 131, 96]. Karamardian [81; 1971] was the first to establish the relationship between the variational inequality problem (P_{VI}) in (4.3) and the complementarity problem (P_{CP}). The existence result for variational inequality problems, the one of theorem 4.1, is taken from [46; theorem 2A.1]. Surveys on variational inequality problems can be found in [63, 50, 76].

Joseph [79; 1979] considers the complementarity problem (P_{CP}), hence a functional inclusion in which the multifunction N is the normal cone to a closed convex cone K , and shows existence, uniqueness, and convergence of a sequence satisfying the iterations (4.12), provided the sought solution x_* is **strongly regular** in the sense of Robinson [120]. Bonnans [19; 1994] introduces the weaker regularity condition that is presented here, namely the semi-stability and hemi-stability of the sought solution, and so provides the strongest local convergence result of the functional inclusion problem (P_{FI}) known so far. The results presented in this section are essentially those in [19]. For an analysis of the local convergence of the *inexact* **JN** algorithm, in which the recurrence formula (4.12) is solved inexactly, we refer the reader to [75]; many algorithms enter that framework, like the stabilized version of SQP and the linearly constrained augmented Lagrangian methods. Joseph has also given a Kantorovich-like existence result based on the iterations (4.12) in [79] and a quasi-Newton analysis of the algorithm in [80]. Analyses of the inexact version of the Joseph-Newton algorithm have been undergone in [75, 45].

Exercises

4.1.1. *Explicit variational problem.* Let \mathbb{E} be a Euclidean space, X be a (not necessarily convex) subset of \mathbb{E} , and $F : \mathbb{E} \rightarrow \mathbb{E}$ be a smooth map. Consider the variational problem

$$(P_v) \quad F(x) + N_X(x) \ni 0,$$

where $N_X(x)$ is the normal cone to X at x . Suppose that \mathbb{F} is another Euclidean space and that X has actually the following form

$$X := \{x \in \mathbb{E} : c(x) \in G\},$$

in which $c : \mathbb{E} \rightarrow \mathbb{F}$ is a smooth function and G is a closed convex set of \mathbb{F} . Show that if x_* is a solution to (P_v) satisfying

$$0 \in \text{int}\{c(x_*) + c'(x_*)\mathbb{E} - G\},$$

then, there exists λ_* in the normal cone to G at $c(x_*)$ such that

$$F(x_*) + c'(x_*)^* \lambda_* = 0.$$

4.1.2. *Equation formulation of a variational inequality problem.* Show that $x \in \mathbb{E}$ is a solution to the *variational inequality problem* (4.3) if and only if x solves (4.4).

4.1.3. *Normal map reformulations.*

1) *Variational inequality problem* [123]. Show that $x \in \mathbb{E}$ solves the variational inequality problem (P_v) in (4.3) if and only if $x = P_C(z)$ where $z \in \mathbb{E}$ solves

$$F(P_C(z)) + z - P_C(z) = 0. \tag{4.36}$$

2) *Complementarity problem.* Show that $x \in \mathbb{E}$ solves the complementarity problem (P_c) in (4.6) if and only if there is a $z \in \mathbb{E}$ such that (x, z) solves

$$F(x) = P_K(z) \quad \text{and} \quad G(x) = -P_{K^-}(z). \tag{4.37}$$

4.1.4. *JN algorithm for a complementarity problem.* Let \mathbb{E} and \mathbb{F} be two Euclidean spaces, K be a nonempty closed convex cone of \mathbb{F} , K^+ its positive dual, and F and $G : \mathbb{E} \rightarrow \mathbb{F}$ be two differentiable functions. Consider the complementarity problem

$$K \ni G(x) \perp F(x) \in K^+. \tag{4.38}$$

Show that the algorithm that computes the next iterate x_{k+1} from the current one x_k by solving the linear complementarity problem in x

$$K \ni \left(G(x_k) + G'(x_k)(x - x_k) \right) \perp \left(F(x_k) + F'(x_k)(x - x_k) \right) \in K^+, \tag{4.39}$$

is the JN algorithm on a certain *functional inclusion problem* like (4.1), in which the multifunction N is the normal cone map to a convex cone; which one?

4.2 B-Newton Method for Systems of Equations ▲

4.3 Linearization method for \mathcal{PC}^1 functions ▲

4.4 Semismooth Newton Method for nonlinear systems

Let \mathbb{E} and \mathbb{F} be two finite dimensional vector spaces. In this section, we present a linearization algorithm for solving a nonlinear system of equations of the form

$$F(x) = 0, \tag{4.40}$$

where $F : \Omega \rightarrow \mathbb{F}$ is a *nonsmooth* function defined on an open set $\Omega \subseteq \mathbb{E}$. The nonsmoothness means here that F is differentiable, but in a weaker sense than that of Fréchet. We shall see that a fast (superlinear or quadratic) local convergence can be obtained if F is *semismooth*, a concept presented in section 4.4.3.

Section 4.4.1 presents some examples where such nonsmooth systems occur. It also discuss an example of function F showing that the Lipschitz continuity of F is not a sufficiently strong assumption for guaranteeing the convergence of the Newton algorithm, since this one may then cycle whatever the solution proximity of the initial iterate can be. A first guess of an appropriate smoothness assumption ensuring the local convergence of a Newton-like algorithm is also presented. Section 4.4.3 defines the concept of semismoothness and gives some of its main properties. A remarkable one is that it is transmissible to the minimum or maximum of two semismooth functions, which makes the concept widespread. The semismooth Newton algorithm is set out in section 4.4.4 and its local convergence is analyzed.

Prerequisite: generalized differentiability (section 4.4.2).

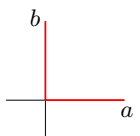
4.4.1 Motivation, Orientation, Examples

Let us start by giving some examples of problems that can be reformulated as nonsmooth systems of equations.

Examples 4.13 1) *Reformulation of a complementarity problem.* Consider the complementarity problem (4.6), in which $\mathbb{F} = \mathbb{R}^n$, $K = \mathbb{R}_+^n$, and F and G are renamed A and B , to avoid confusion with the function F introduced above:

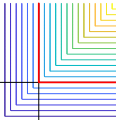
$$0 \leq A(x) \perp B(x) \geq 0. \tag{4.41}$$

In this system, A and $B : \mathbb{E} \rightarrow \mathbb{R}^n$ are two functions defined on a vector space \mathbb{E} . This problem means that one seeks a point $x \in \mathbb{E}$ such that $A(x) \geq 0$, $B(x) \geq 0$, and $A(x)^\top B(x) = 0$ (or equivalently $A_i(x)B_i(x) = 0$ for all $i \in [1:n]$, which highlights the combinatorial aspect of the problem). Such a problem can be written in the form of a nonsmooth equation, thanks to a C-function. A *C-function* (C for complementarity) is a function $\varphi : \mathbb{R}^2 \rightarrow \mathbb{R}$ such that

$$\varphi(a, b) = 0 \iff a \geq 0, \quad b \geq 0, \quad ab = 0. \tag{4.42}$$


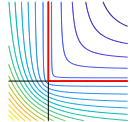
Hence $\varphi(a, b) = 0$ expresses the complementarity of the two scalars a and b . The most frequently encountered C-functions are the C-function “min” and the Fischer C-function.

- The *min C-function* is defined by

$$\varphi(a, b) = \min(a, b). \tag{4.43}$$


It is easy to verify that (4.42) holds for this function. The min function is concave (minimum of two linear functions) and nondifferentiable at points (a, b) verifying $a = b$.

- The *Fischer C-function* is defined by

$$\varphi(a, b) = \sqrt{a^2 + b^2} - (a + b). \tag{4.44}$$


It is also easy to verify that (4.42) holds for this function. This function is convex (it is the ℓ_2 norm of (a, b) minus a linear function) and differentiable everywhere except at $(0, 0)$.

With a C-function φ , the complementarity problem (4.41) can be written like the nonsmooth equation

$$F(x) \equiv \begin{pmatrix} \varphi(A_1(x), B_1(x)) \\ \vdots \\ \varphi(A_n(x), B_n(x)) \end{pmatrix} = 0.$$

- 2) *Reformulation of a variational inequality problem.* The variational inequality problem (4.3), written here $\Phi(x) + N_C(x) \ni 0$, where N_C is the multifunction “normal cone to the nonempty closed convex set C ”, can be rewritten like a nonsmooth equation $F(x) = 0$ by observing that $-\Phi(x) \in N_C(x)$ if and only if the projection of $x - \Phi(x)$ on C is x [47]. Then, it suffices to define $F : \mathbb{E} \rightarrow \mathbb{E}$ at $x \in \mathbb{E}$ by

$$F(x) := x - P_C(x - \Phi(x)),$$

where P_C is the orthogonal projector on C . Note that the boundary of C must have some smoothness to make this approach work (recall remark 1.12).

Another way of reformulating a variational inequality problem as a nonsmooth equation is to introduce an equation characterizing the point $z := x - \Phi(x)$ instead of x . This point of view is examined in the exercise 4.1.3.

Below, we consider the generalization of the Newton algorithm for solving the nonsmooth equation (4.40) that reads locally

$$x_{k+1} := x_k - J_k^{-1}F(x_k), \tag{4.45}$$

where $J_k \in \mathcal{L}(\mathbb{E}, \mathbb{F})$ is a nonsingular element of the Clarke differential $\partial_C F(x_k)$. The target we fix ourselves is very ambitious, since the iteration (4.45) only requires the solution to a linear system, which is much less expensive than the JN algorithm, whose iteration is potentially nonlinear. We shall see that one can take assumptions on F , often verified, such that locally this algorithm generates sequences that converge superlinearly or quadratically to a “regular” zero of F . Therefore, one recovers the two conditions that are required for ensuring the convergence of the Newton algorithm: a certain smoothness of F and the regularity of the sought zero, in a sense that still needs to be clarified.

Let us start by observing that, without an adequate smoothness assumption on F , the proposed generalization (4.45) of Newton’s method is doomed for failure, even though the iterates only visit points of differentiability of F , with nonsingular Jacobians. This is what is shown in the next example.

Counter-example 4.14 (nonconvergence of the Newton algorithm for a nonsmooth equation [88]) We construct a function $F : [-1, 1] \subseteq \mathbb{R} \rightarrow [-1, 1] \subseteq \mathbb{R}$, by repeating indefinitely a pattern that is scaled in proportion to its proximity to the unique zero $x_* = 0$ of the function. The pattern is designed in order to force the Newton method to make a cycle, which can be arbitrary close to the zero (the graph of the function is given on the right-hand side of figure 4.1).

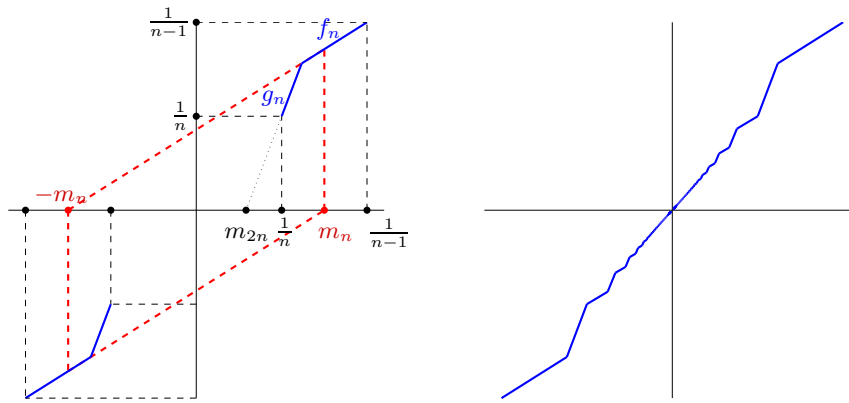


Fig. 4.1. Kummer’s function [88], for which the Newton method makes cycles, as close as desired to its unique zero $x_* = 0$. The function is constructed by repeating the pattern on the left-hand side, scaling it in proportion to its proximity to the zero.

Here is how this function is constructed. The function is defined on $[-1, 1]$, is continuous and odd (meaning that $F(-x) = -F(x)$ for all $x \in [-1, 1]$), so that it suffices to define it on $(0, 1]$ and to set $F(0) = 0$. This is done by means of a function pattern that is piecewise affine on the intervals $[1/n, 1/(n-1)]$ for all integer $n \geq 2$, takes the value $1/(n-1)$ at $x = 1/(n-1)$ and the value $1/n$ at $x = 1/n$ (see the graph of the pattern in the left-hand side of figure 4.1). This pattern is then reproduced and scaled to define the function on $(0, 1]$. Let us denote by

$$m_n := \frac{1}{2} \left(\frac{1}{n} + \frac{1}{n-1} \right) = \frac{2n-1}{2n(n-1)}$$

the middle point of the interval $[1/n, 1/(n-1)]$. This point will be a point of a cycle for Newton's method, if F is affine near m_n with the values

$$f_n(x) = a_n(x + m_n), \quad \text{where } a_n = \frac{1/(n-1)}{1/(n-1) + m_n} = \frac{2n}{4n-1}.$$

Indeed, in this case, the iterate that follows m_n is $-m_n$ (because $f_n(-m_n) = 0$), itself followed by m_n (by the oddness of the function); hence the Newton algorithm (4.45) cycles between m_n and $-m_n$. The second affine part of the pattern is the function g_n that must take the value $1/n$ at $x = 1/n$ (so that the connexion with the following pattern is done continuously) and must have a sufficiently large slope b_n (so that its intersection with f_n occurs at an abscissa lower than m_n (so that the previously mentioned cycle can still occur). As we shall see, this last condition is satisfied if g_n vanishes at $x = m_{2n}$, i.e., if

$$g_n(x) = b_n(x - m_{2n}), \quad \text{with } b_n = \frac{1/n}{1/n - m_{2n}} = \frac{8n-4}{4n-3}.$$

To check that the slope b_n of g_n is large enough, it suffices to verify that, for any integer $n \geq 2$, we have $f_n(m_n) < g_n(m_n)$, which reads $4n^2 + n - 1 > 0$, an inequality that is verified for all positive integers.

The function F so obtained has its graph represented on the right-hand side in figure 4.1. Its unique zero is $x_* = 0$. It is Lipschitz continuous on $[-1, 1]$, with the modulus $\max_n b_n = b_2 = 12/5$. Since, for all x in the interval $[1/n, 1/(n-1)]$, $F(x)$ is in the same interval, it follows that $F(x)/x \in [(n-1)/n, n/(n-1)]$ and therefore, by parity, $F'(0) = 1$. The C-differential $\partial_C F(0)$ is the convex hull of the limits of the a_n and b_n when $n \rightarrow \infty$, by the definition 4.15, that is $[1/2, 2]$. This C-differential does not contain the zero slope. In summary:

$$F : \mathbb{R} \rightarrow \mathbb{R} \text{ is Lipschitz continuous and has directional derivatives,} \quad (4.46a)$$

$$F(0) = 0, \quad (4.46b)$$

$$F'(0) = 1, \quad (4.46c)$$

$$\partial_C F(0) = [\tfrac{1}{2}, 2] \neq 0. \quad (4.46d)$$

□

What is wrong with the nonsmooth function F in example 4.14? Is there no hope to get local convergence of the Newton algorithm with a nonsmooth function? An answer is given by the following expression of the error $x_{k+1} - x_*$:

$$\begin{aligned} x_{k+1} - x_* &= x_k - x_* - J_k^{-1} F(x_k) && [(4.45)] \\ &= -J_k^{-1} [F(x_k) - F(x_*) - J_k(x_k - x_*)] && [F(x_*) = 0] \\ &= -J_k^{-1} [F(x_* + h_k) - F(x_*) - J_k h_k] && [h_k := x_k - x_*]. \end{aligned}$$

As a result, if $\{J_k^{-1}\}$ is bounded and if

$$F(x_* + h_k) - F(x_*) - J_k h_k = o(\|h_k\|), \tag{4.47}$$

the superlinear convergence of the generated sequence is guaranteed. Conversely, if $\{J_k\}$ is bounded and the generated sequence converges superlinearly, then (4.47) holds. Hence this condition is almost necessary and sufficient to get superlinear convergence.

The condition (4.47) makes no assumption on the way of choosing J_k (only the boundedness of $\{J_k\}$ and $\{J_k^{-1}\}$ is assumed in the discussion). In the semismooth Newton algorithm of section 4.4.4 below, J_k is chosen in $\partial_C F(x_k)$, so that (4.47) becomes a condition of the generalized differential $\partial_C F(x_k) = \partial_C F(x_* + h_k)$, not a condition on $\partial_C F(x_*)$, as an analogy with the Fréchet differentiability condition “ $F(x_* + h_k) - F(x_*) - F'(x_*)h_k = o(\|h_k\|)$ ” would incline us to do. This small change is a fundamental aspect of semismoothness, which takes indeed the condition (4.47) in its definition 4.20. We shall see that this notion of weak differentiability is to be shared by a large number of functions, so that superlinear convergence of a Newton-like method is attainable for many nonsmooth systems of equations. This is good news.

Let us come back to the counter-example 4.14 and show that the condition (4.47) does not hold at $x_* = 0$, which explains a posteriori why the local convergence of the Newton method cannot be guaranteed for the constructed function. The Jacobians J_k are the slopes a_n , whose inverses $a_n^{-1} = 2 - 1/(2n)$ form a bounded sequence. Consider now the condition (4.47). If we take a sequence of points $h_n \in (1/n, y_n)$, where y_n is the abscissa of the intersection of f_n and g_n , we have

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{|F(0 + h_n) - F(0) - F'(h_n) \cdot h_n|}{|h_n|} &= \lim_{n \rightarrow \infty} \frac{|b_n(h_n - m_{2n}) - b_n h_n|}{h_n} \\ &\geq \lim_{n \rightarrow \infty} \frac{b_n m_{2n}}{m_n} \quad [h_n \leq m_n] \\ &= 1, \end{aligned}$$

which is nonzero. Taking $h_n = m_n$ would have given the limit 1/2. Hence, the condition (4.47) does not hold.

4.4.2 Generalized Differentiability

Definition

Let \mathbb{E} and \mathbb{F} be two finite dimensional normed spaces and $F : \mathbb{E} \rightarrow \mathbb{F}$ be a function. We assume that the reader is aware of the notions introduced in sections 1.3.2 and 1.3.4.

Definitions 4.15 (B-differential, C-differential) The *B-differential*³ of F at x is the set denoted and defined by

$$\partial_B F(x) := \{J \in \mathcal{L}(\mathbb{E}, \mathbb{F}) : \exists \{x_k\} \subseteq \mathcal{D}_F \text{ such that } x_k \rightarrow x, F'(x_k) \rightarrow J\}.$$

The *C-differential*⁴ of F at x is the convex hull of the B-differential, namely

$$\partial_C F(x) := \text{co } \partial_B F(x). \quad \square$$

³ The “B” of the B-differential is honoring Bouligand [122].

⁴ The “C” of the C-differential is for Clarke.

If $F : \mathbb{E} \rightarrow \mathbb{F}$ is differentiable near $x \in \mathbb{E}$ and if F' is continuous at x , it is clear that $\partial_B F(x) = \partial_C F(x) = \{F'(x)\}$. But if F is Lipschitz near x and only differentiable at x , $\partial_C F(x)$ is not necessarily a singleton (Kummer's function in section 4.4.1, with its properties (4.46c) and (4.46d), will provide an example of this curiosity); nevertheless, by taking $x_k = x$ for all k in the definition of $\partial_B F(x)$, we see that $F'(x) \in \partial_B F(x)$ in this case.

We refer the reader to the monograph [32] for a detailed study of the C-differential. Below, we try to say as little as possible, although enough to cover our future needs. The notion of upper semi-continuous multifunction used below has been introduced in section 1.3.4.

Proposition 4.16 (compactness and upper semi-continuity) *If $F : \Omega \rightarrow \mathbb{F}$ is L -Lipschitz near $x \in \Omega$, then*

- 1) $\partial_C F(x)$ is nonempty compact ($\subseteq L\bar{B}$) and convex,
- 2) $\partial_C F$ is upper semi-continuous at x .

Proof. 0) Let us start by showing that $\|F'(x_0)\| \leq L$ for all $x_0 \in \mathcal{D}_F$ near x . By the differentiability of F at x_0 , it follows that $F(x') = F(x_0) + F'(x_0)(x' - x_0) + o(\|x' - x_0\|)$. Hence, for small $\varepsilon > 0$, there exists $\delta > 0$ such that, for $x' \in \bar{B}(x_0, \delta)$, one has $\|F'(x_0)(x' - x_0)\| \leq \|F(x') - F(x_0)\| + \varepsilon\|x' - x_0\| \leq (L + \varepsilon)\|x' - x_0\|$. We deduce from this that $\|F'(x_0)\| \leq L + \varepsilon$, and next that $\|F'(x_0)\| \leq L$ since $\varepsilon > 0$ is arbitrary.

1) The C-differential is convex by construction. Next, it suffices to show that $\partial_B F(x)$ is nonempty, closed, and in $L\bar{B}$ (since, according to (1.2), the convex hull of a compact set is compact).

- [$\partial_B F(x) \neq \emptyset$] Since $\Omega \setminus \mathcal{D}_F$ has measure zero, one can find a sequence $\{x_k\} \subseteq \mathcal{D}_F$ converging to x (otherwise there would exist an $\varepsilon > 0$ such that $x + \varepsilon B \subseteq \Omega \setminus \mathcal{D}_F$, which is in contradiction with the zero measure of $\Omega \setminus \mathcal{D}_F$). Since the sequence $\{F'(x_k)\}$ is bounded (point 0), one can extract a convergent subsequence. The limit of this one is therefore in $\partial_B F(x)$, by definition of this last set.
- [$\partial_B F(x)$ is closed] Indeed, if $\{J_k\} \subseteq \partial_B F(x)$ and $J_k \rightarrow J$, then, for all $\varepsilon > 0$, one can find an index k such that $\|J_k - J\| \leq \varepsilon$ and a point $x_k \in \mathcal{D}_F$ such that $\|F'(x_k) - J_k\| \leq \varepsilon$ and $\|x_k - x\| \leq \varepsilon$. Clearly, when $\varepsilon \downarrow 0$, the sequence $\{x_k\}$ so constructed converges to x and $F'(x_k)$ converges to J . Hence $J \in \partial_B F(x)$.
- [$\partial_B F(x) \subseteq L\bar{B}$] Any element $J \in \partial_C F(x)$ is the limit of operators $F'(x_k)$ with a norm not exceeding L (point 0). Hence $\|J\| \leq L$, by the continuity of the norm.

2) It suffices to show that,

$$\forall \varepsilon > 0, \exists \delta > 0, \forall x' \in x + \delta B : \partial_B F(x') \subseteq \partial_C F(x) + \varepsilon B.$$

Indeed, this last inclusion implies the desired inclusion $\partial_C F(x') \subseteq \partial_C F(x) + \varepsilon B$, since $\partial_C F(x) + \varepsilon B$ is a convex set.

We proceed by contradiction. If the claim is not true, there exists $\varepsilon > 0$ and a sequence $\{x_k\} \subseteq \mathbb{E}$ converging to x and elements $J_k \in \partial_B F(x_k)$ that are not in $\partial_B F(x) + \varepsilon B$. Then,

$$(J_k + \frac{\varepsilon}{2}B) \cap (\partial_B F(x) + \frac{\varepsilon}{2}B) = \emptyset.$$

The fact that $J_k \in \partial_B F(x_k)$ implies, by definition, that there is a point $\tilde{x}_k \in \mathcal{D}_F$ such that $\|\tilde{x}_k - x_k\| \leq 1/k$ and $F'(\tilde{x}_k) \in J_k + \frac{\epsilon}{2}B$. On the other hand, since $x_k \rightarrow x$, the sequence $\{F'(x_k)\}$ is bounded and, by extracting a subsequence if needed, one can assume that it converges to some J . Then, one would have $\tilde{x}_k \rightarrow x$ and $F'(\tilde{x}_k) \rightarrow J$, so that $J \in \partial_B F(x)$. But then $F'(\tilde{x}_k) \in \partial_B F(x) + \frac{\epsilon}{2}B$ for k large. We have the contradiction since $F'(\tilde{x}_k)$ belongs to the two sets $J_k + \frac{\epsilon}{2}B$ and $\partial_B F(x) + \frac{\epsilon}{2}B$, which have been shown to be disjoint. \square

Properties

The standard quadratic convergence result of the Newton method to solve the smooth system $F(x) = 0$ requires that the Jacobian $F'(x_*)$ be nonsingular at the sought root x_* (this has been recalled in proposition 1.48). When $F \in \mathcal{C}^1$, this nonsingularity property is “diffused” to the Jacobians $F'(x)$ if x is close to x_* (this is due to the continuity of F' and the fact that the set of nonsingular operators is open), which makes the Newton algorithm well defined in a neighborhood of a “regular” root of F . This diffusion property is made precise in the following Banach perturbation lemma, which is recalled below in the finite dimension setting.

Lemma 4.17 (Banach perturbation lemma) *Let \mathbb{E} and \mathbb{F} be two vector spaces, $A : \mathbb{E} \rightarrow \mathbb{F}$ be a nonsingular linear operator, and $B : \mathbb{E} \rightarrow \mathbb{F}$ be another linear operator sufficiently close to A in the sense that $\|A^{-1}(B - A)\| < 1$. Then, B is also nonsingular and*

$$\|B^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}(B - A)\|}.$$

In the case of the semismooth Newton algorithm of section 4.4.4, this regularity assumption becomes the nonsingularity of all the generalized Jacobians in $\partial_C F(x_*)$, a property that is called C-regularity.

Definitions 4.18 (regular C-differential, C-regular point) The C-differential $\partial_C F(x)$ is said to be *regular* if all its jacobians $J \in \partial_C F(x)$ are nonsingular. Then, one also say that the point x is *C-regular* for F . \square

The next proposition shows a property similar to the Banach perturbation lemma, but for a locally Lipschitz function F : the points close to a C-regular point for F are also C-regular and a bound on the Jacobians of their C-differential, and their inverse, can be given.

Proposition 4.19 (C-regularity diffusion) *If $F : \mathbb{E} \rightarrow \mathbb{F}$ is Lipschitz near a C-regular point $x \in \mathbb{E}$, then, there are constants $C > 0$ and $\delta > 0$ such that any*

point in $B(x, \delta)$ is C -regular and

$$\sup_{\substack{x' \in B(x, \delta) \\ J \in \partial_C F(x')}} \max (\|J\|, \|J^{-1}\|) \leq C. \quad (4.48)$$

Proof. 1) Let us first show the bound on $\|J\|$. By point 1 of proposition 4.16, we know that $\partial_C F(x) \subseteq L_x \bar{B}$, for a constant $L_x \geq 0$ depending on x . We now use the upper semi-continuity of $\partial_C F$ with the open neighborhood $\partial_C F(x) + B$ of $\partial_C F(x)$: one can find $\varepsilon > 0$ such that, if $x' \in B(x, \varepsilon)$, it follows that $\partial_C F(x') \subseteq \partial_C F(x) + B$. As a result, for all $x' \in B(x, \varepsilon)$ and all $J \in \partial_C F(x')$, we have $\|J\| \leq L_x + 1$.

2) Let us now show the bound on $\|J^{-1}\|$. Consider first a particular Jacobian $J_0 \in \partial_C F(x)$. This one is nonsingular by the C -regularity of x . By the Banach perturbation lemma, if $\|J - J_0\| \leq \varepsilon(J_0) := 1/(2\|J_0^{-1}\|)$, it follows that $\|J_0^{-1}(J - J_0)\| \leq 1/2$ and therefore

$$\|J^{-1}\| \leq \frac{\|J_0^{-1}\|}{1 - \|J_0^{-1}(J - J_0)\|} \leq 2\|J_0^{-1}\|. \quad (4.49)$$

Consider now a cover of $\partial_C F(x)$ by the open sets $\{J + \varepsilon(J)B : J \in \partial_C F(x)\}$. By the compactness of $\partial_C F(x)$ (point 1 of proposition 4.16), one can extract a finite sub-cover (Heine-Borel-Lebesgue property): one can find m elements $J_i \in \partial_C F(x)$, such that the open set

$$V := \bigcup_{i \in [1 : m]} (J_i + \varepsilon(J_i)B)$$

covers $\partial_C F(x)$. Remark that the operators in V have bounded inverses: for all $J \in V$, one has

$$\|J^{-1}\| \leq \max_{i \in [1 : m]} 2\|J_i^{-1}\| =: \beta. \quad (4.50)$$

thanks to (4.49). It now suffices to show that $\partial_C F(x') \subseteq V$ when x' is close to x . To this end, one uses the compactness of $\partial_C F(x)$ and the upper semi-continuity of $\partial_C F$.

By compactness of $\partial_C F(x)$, one can find an $\varepsilon > 0$ such that

$$\partial_C F(x) + \varepsilon B \subseteq V.$$

Indeed, otherwise, one could find operator sequences $\{J'_k\}$ and $\{J''_k\}$ such that $J'_k \notin V$, $J''_k \in \partial_C F(x)$, and $\|J'_k - J''_k\| \leq 1/k$. By compactness $\partial_C F(x)$, one can extract a subsequence of $\{J''_k\}$ converging to some $J \in \partial_C F(x)$. Obviously $J'_k \rightarrow J$, so that J'_k is in some $J_i + \varepsilon(J_i)B$ for some $i \in [1 : m]$, and therefore $J'_k \in V$, which contradicts the starting assumption.

Finally, by the upper semi-continuity of $\partial_C F$: one can find a $\delta > 0$ such that, if $x' \in x + \delta B$, one has $\partial_C F(x') \subseteq \partial_C F(x) + \varepsilon B \subseteq V$, which is what we wanted to show. □

4.4.3 Semismoothness \blacktriangle

Definition

Throughout this section, \mathbb{E} and \mathbb{F} are two normed spaces and Ω be an open set of \mathbb{E} .

Definitions 4.20 (semismoothness) Let $F : \Omega \rightarrow \mathbb{F}$ be a function and $x \in \Omega$. The function F is said to be *semismooth* at x if the following three conditions hold:

- (SS1) F is Lipschitz near x ,
- (SS2) F has directional derivatives at x in all directions,
- (SS3) when $h \rightarrow 0$ in \mathbb{E} , one has

$$\sup_{J \in \partial_C F(x+h)} \|F(x+h) - F(x) - Jh\| = o(\|h\|). \quad (4.51a)$$

The function F is said to be *strongly semismooth* at x if it is strongly semismooth at x with (SS3) strengthened into

- (SS3') for h near 0, one has

$$\sup_{J \in \partial_C F(x+h)} \|F(x+h) - F(x) - Jh\| = O(\|h\|^2). \quad (4.51b)$$

The function $F : \Omega \rightarrow \mathbb{F}$ is said to be *semismooth* (resp. *strongly semismooth*) on a part P of Ω if it is semismooth (resp. strongly semismooth) at all points of P . \square

- Remarks 4.21** 1) The local Lipschitz continuity of F at x in (SS1) guarantees that the Clarke differential $\partial_C F$ is well defined and nonempty near x (point 1 of proposition 4.16), so that its use in (SS3) and (SS3') makes sense. Property (SS2) is useful so that the semismoothness enjoys good properties, those given in propositions 4.23-4.29 below. As announced in the discussion of section 4.4.1, property (SS3) is the one that ensures the superlinear convergence of the semismooth Newton algorithm and property (SS3') its quadratic convergence (theorem 4.31). We stress again the fact that, in (SS3) and (SS3'), the Jacobians J are taken in the C-differential $\partial_C F(x+h)$ and not in $\partial_C F(x)$, as one could be tempted for mimicking the Fréchet differentiability; the motivation of this choice has been given in section 4.4.1.
- 2) The differential $\partial_C F$ is not always easy to compute. However, if an overestimate D of this one is known, in the sense that $\partial_C F(x') \subseteq D(x')$ for all x' near the point x of interest, and if (4.51a) or (4.51b) can be verified with $\partial_C F$ replaced by D , the corresponding smoothness property will hold.

As shown in the next proposition, semismoothness (resp. strong semismoothness) can be defined by other properties. More precisely, the assumption (SS3) (resp. (SS3')) can be replaced by other assumptions.

Proposition 4.22 (semismoothness characterizations) *Let $F : \Omega \rightarrow \mathbb{F}$ be a function and $x \in \Omega$. Suppose that F satisfies (SS1) and (SS2). Then the following properties are equivalent:*

- (i) F is semismooth (resp. strongly semismooth) at x ,
- (ii) for $h \rightarrow 0$, there holds

$$\sup_{J \in \partial_C F(x+h)} \|Jh - F'(x; h)\| = o(\|h\|) \quad (\text{resp. } = O(\|h\|^2)),$$
- (iii) for $h \rightarrow 0$ such that $x + h \in \mathcal{D}_F$, there holds

$$F'(x+h)h - F'(x; h) = o(\|h\|) \quad (\text{resp. } = O(\|h\|^2)).$$

Proof. □

Properties

Here are a few useful properties of semismooth functions. There is no need to know them for proving theorem 4.31 giving the local convergence of the semismooth Newton algorithm, but they are very useful for recognizing the semismoothness of functions and therefore for knowing whether the convergence result applies to them.

The first property tells us that sufficiently smooth functions are semismooth, which is somehow reassuring.

Proposition 4.23 (differentiability and semismoothness) *If $F : \Omega \rightarrow \mathbb{F}$ is differentiable near $x \in \Omega$ and F' is continuous at x (resp. radially Lipschitz at x), then F is semismooth (resp. strongly semismooth) at x .*

Proof. [(SS1)] By the mean value theorem 1.24, for x_1 and x_2 near x , there holds

$$\|F(x_2) - F(x_1)\| \leq \left(\sup_{z \in (x_1, x_2)} \|F'(z)\| \right) \|x_2 - x_1\|.$$

By the continuity of F' at x , the first factor in the right-hand side is as close to $\|F'(x)\|$ as desired, by taking x_1 and x_2 sufficiently close to x . Le Lipschitz continuity of F near x follows.

[(SS2)] This property is a consequence of the differentiability of F at x : $F'(x; h) = F'(x)h$.

[(SS3)] By the differentiability of F at x and the continuity of F' at x , we get

$$F(x+h) - F(x) - F'(x+h)h = (F'(x) - F'(x+h))h + o(\|h\|) = o(\|h\|).$$

[(SS3')] Assume now that F' is radially L-Lipschitz at x . By the mean value corollary 1.25, for h small, there holds

$$\begin{aligned}
& \|F(x+h) - F(x) - F'(x+h)h\| \\
& \leq \left(\sup_{z \in (x, x+h)} \|F'(z) - F'(x+h)\| \right) \|h\| \\
& \leq L \left(\sup_{t \in (0,1)} \|(1-t)x + t(x+h) - (x+h)\| \right) \|h\| \\
& \leq L\|h\|^2.
\end{aligned}$$

This shows the strong semismoothness of F at x . \square

Proposition 4.24 (semismoothness of a convex function) *If $f : \Omega \rightarrow \mathbb{R}$ is convex on the open convex set $\Omega \subseteq \mathbb{E}$ and if $x \in \Omega$, then f is semismooth at x .*

Proof. A convex function is Locally Lipschitz on the relative interior of its domain [65], hence certainly on Ω in our case. It has also directional derivatives at any point of its domain (but these can have infinite values), which have here finite values since f takes only finite values on Ω . In addition, $\partial_C f(x)$ is the subdifferential of f at x in the sense of convex analysis [32; propositions 2.2.6 and 2.2.7]. . . . (for the sequel, see [100] or the prof of [50; prop. 7.4.5(c)]). \square

In the proposition below, $F : \Omega \rightarrow \mathbb{F}$ is said to be *piecewise semismooth* at $x \in \Omega$, if there is a neighborhood V of x and semismooth functions $F_i : V \rightarrow \mathbb{F}$, for $i \in I$ with finite $|I|$, such that, for all $x' \in V$, there is an index $i \in I$ such that $F(x') = F_i(x')$.

Proposition 4.25 (piecewise semismoothness) *If $F : \Omega \rightarrow \mathbb{F}$ is piecewise semismooth at $x \in \Omega$, then F is semismooth en x .*

The function F is said to be *piecewise affine* at x if the pieces in the definition of piecewise semismoothness are affine.

Proposition 4.26 (piecewise affinity) *If $F : \Omega \rightarrow \mathbb{F}$ is piecewise affine at $x \in \Omega$, then F is strongly semismooth at x .*

The semismoothness of a vector-valued function can be deduced from the semismoothness of its components.

Proposition 4.27 (componentwise semismoothness) *Let $F_1 : \Omega \rightarrow \mathbb{F}_1$ and $F_2 : \Omega \rightarrow \mathbb{F}_2$ be two functions with values in the finite dimensional normed spaces \mathbb{F}_1 and \mathbb{F}_2 , and let $x \in \Omega$. Then,*

$$(F_1, F_2) : x' \in \Omega \rightarrow (F_1(x'), F_2(x')) \in \mathbb{F}_1 \times \mathbb{F}_2$$

is semismooth (resp. strongly semismooth) at x if and only if F_1 and F_2 are semismooth (resp. strongly semismooth) at x .

Semismoothness is stable by composition.

Proposition 4.28 (semismoothness and composition) *If $F : \Omega \rightarrow \mathbb{F}$ is semismooth (resp. strongly semismooth) at $x \in \Omega$, if $\Omega_{\mathbb{F}}$ is a neighborhood of $F(x)$ in \mathbb{F} , and if $G : \Omega_{\mathbb{F}} \rightarrow \mathbb{G}$ is semismooth (resp. strongly semismooth) at $F(x)$, then $G \circ F$ is semismooth (resp. strongly semismooth) at x .*

An important asset of the semismoothness is to be stable with respect to the minimum or maximum of functions, which is of course not the case for the Fréchet differentiability! Since many nonsmooth functions can be defined by using these operators, these functions are often semismooth.

Proposition 4.29 (calculus) *If $F_1 : \Omega \rightarrow \mathbb{F}$ and $F_2 : \Omega \rightarrow \mathbb{F}$ are semismooth (resp. strongly semismooth) at x , then the following functions are semismooth (resp. strongly semismooth) at x (for the last two, $\mathbb{F} = \mathbb{R}^m$ and the operators “max” and “min” act componentwise):*

$$F_1 + F_2, \quad \langle F_1, F_2 \rangle, \quad \max(F_1, F_2), \quad \text{and} \quad \min(F_1, F_2).$$

Proof. The semismoothness result for $F_1 + F_2$ can be obtained by viewing this function as the composition of $x \mapsto (F_1(x), F_2(x))$, which is semismooth (resp. strongly semismooth) by proposition 4.27, and the addition $(u, v) \mapsto u + v$, that is linear. Now, use proposition 4.28 to conclude.

The function $\langle F_1, F_2 \rangle$ is also a composition of semismooth (resp. strongly semismooth) functions, namely $x \mapsto (F_1(x), F_2(x))$ like above and the scalar product of \mathbb{F} , which is bilinear, hence C^∞ .

The function $\max(F_1, F_2)$ (the same reasoning holds for $\min(F_1, F_2)$) is still the composition of $x \mapsto (F_1(x), F_2(x))$ and the function $(u, v) \in \mathbb{F} \times \mathbb{F} \mapsto \max(u, v)$ which is piecewise linear, hence strongly semismooth by proposition 4.26. \square

Here are a few examples and counter-examples of (strongly) semismooth functions.

- Examples 4.30** 1) We already know that a norm is semismooth, due to its convexity (proposition 4.24). The ℓ_p norms are, furthermore, *strongly* semismooth (exercise 4.4.1).
 2) The “min” and Fischer C-functions, (4.43) and (4.44), are strongly semismooth (exercise 4.4.2).

- 3) The projector on a convex set defined by \mathcal{C}^2 constraints is strongly semismooth at a point of the convex set that satisfies (CQ-LI) (exercise ??).
- 4) For an arbitrary convex set C , the projector P_C may not have directional derivatives at a point not belonging to C [86, 128] and is therefore not semismooth at such a point. This is not good news for solving variational inequality problem using the reformulation in example 4.13(2). Nevertheless, the situation is more favorable if C is a polyhedron or has a smooth boundary. Also semismooth-Newton-like algorithms give excellent results by using some kind of pseudo-generalized Jacobians [93; 2018]. □

4.4.4 The Semismooth Newton Method ▲

The *semismooth Newton algorithm* is designed to solve the system (4.40), in which $F : \Omega \rightarrow \mathbb{F}$ is semismooth (section 4.4.3). Locally (i.e., near a solution x_* to this system), it consists in generating a sequence $\{x_k\}$ in the open set $\Omega \subseteq \mathbb{E}$ by the recurrence

$$x_{k+1} := x_k - J_k^{-1}F(x_k), \tag{4.52}$$

where J_k is a nonsingular Jacobian of the Clarke differential $\partial_C F(x_k)$. To be well defined, it is clear that such a Jacobian must exist in the C-differential of F at the visited points. But the technique of proof used below requires a little more than that, namely the boundedness of the sequence of inverses $\{J_k^{-1}\}$ and the property (SS3) or (SS3') of the semismoothness. In the terms of theorem 4.31 below, the boundedness of $\{J_k^{-1}\}$ is ensured by the regularity of $\partial_C F(x_*)$ and proposition 4.19.

It is important to observe that the semismooth Newton method is particularly computationally sober, since it only requires to solve a linear system per iteration, like the Newton algorithm for smooth systems, but unlike the JN algorithm for functional inclusions (section 4.1.2) or the SQP algorithm for nonlinear optimization (section 5.1), derived from the latter. Now, all these algorithms are not used to solve similar problems and cannot be globalized with the same ease.

Theorem 4.31 (local convergence of the semismooth Newton algorithm) *Suppose that F is semismooth at a C-regular solution x_* to (4.40). Then, there exists a neighborhood V of x_* such that, if the first iterate x_1 is in V , the semismooth Newton algorithm (4.52) is well defined and generates a sequence $\{x_k\}$ in V , which converges superlinearly to x_* (and quadratically if F is strongly semismooth at x_*).*

Proof. The regularity of $\partial_C F(x_*)$ and proposition 4.19 imply that there exist constants $C_1 > 0$ and $\varepsilon_1 > 0$ such that

$$\sup_{\substack{x \in \bar{B}(x_*, \varepsilon_1) \\ J \in \partial_C F(x)}} \max (\|J\|, \|J^{-1}\|) \leq C_1. \tag{4.53}$$

By the semismoothness property (SS3) of F at x_* , one can find $\varepsilon \in (0, \varepsilon_1]$ such that

$$\|F(x) - F(x_*) - J(x - x_*)\| \leq \frac{1}{2C_1} \|x - x_*\|, \quad (4.54)$$

when $x \in \bar{B}(x_*, \varepsilon)$ and $J \in \partial_C F(x)$.

Suppose now that $x_k \in V := \bar{B}(x_*, \varepsilon)$. Then the algorithm is well defined at this point since any Jacobian J_k chosen in $\partial_C F(x_k)$ is nonsingular (by $\varepsilon \leq \varepsilon_1$ and (4.53)). A new iterate x_{k+1} can therefore be computed by (4.52). One has

$$x_{k+1} - x_* = x_k - x_* - J_k^{-1} F(x_k) = -J_k^{-1} (F(x_k) - F(x_*) - J_k(x_k - x_*)),$$

hence

$$\|x_{k+1} - x_*\| \leq \|J_k^{-1}\| \|F(x_k) - F(x_*) - J_k(x_k - x_*)\|. \quad (4.55)$$

By (4.53) and (4.54), $\|x_{k+1} - x_*\| \leq \frac{1}{2} \|x_k - x_*\|$, which shows that $x_{k+1} \in \bar{B}(x_*, \varepsilon) = V$, the wellposedness of the algorithm, and the convergence of the convergence $\{x_k\}$ to x_* .

Then, the semismoothness implies that $F(x_k) - F(x_*) - J_k(x_k - x_*) = o(\|x_k - x_*\|)$, so that $x_{k+1} - x_* = o(\|x_k - x_*\|)$ by (4.55) and (4.53). This is the superlinear convergence of $\{x_k\}$ to x_* .

In case of strong semismoothness, $F(x_k) - F(x_*) - J_k(x_k - x_*) = O(\|x_k - x_*\|^2)$, so that $x_{k+1} - x_* = O(\|x_k - x_*\|^2)$ by (4.55) and (4.53), showing that the convergence is quadratic. \square

4.4.5 Globalization by linesearch \blacktriangle

It is well known that the Newton direction is a descent direction to the least-squares merit function. If this surprising local property is not able to ensure global convergence of the damped Newton iterates to a zero of F , when this one exists, it paves the way to algorithmic techniques having interesting properties, in particular when trust regions are used.

This descent property is no longer guaranteed when F is only semismooth. Nevertheless, when the least-squares merit function is smooth, some interesting properties can be obtained [78; 1999]. We describe this particular case in this section.

Let us start by giving a concrete example of such a situation, which has numerous applications.

Here is a short review of what has been explored.

- Jiang and Ralph [78; 1999] analyze semismooth Newton and Gauss-Newton algorithms, globalized with linesearch or trust regions on the least-squares merit function, provided this one is smooth (with application to the nonlinear complementarity problem).
- Ueda and Yamashita [134; 2012] also suppose that the least-squares merit function is smooth and analyze the complexity of the Levenberg-Morrison-Marquardt approach for nonsmooth equations (with application to the nonlinear complementarity problem).
- See also [114; 2016] (with application to the nonlinear complementarity problem) and many other papers.

4.4.6 Globalization by trust regions ▲**4.4.7 Examples of use ▲****Notes**

The notion of semismoothness was first introduced by Mifflin [100; 1977] for real-valued functions and extended to vector-valued functions by Qi and Sun [112, 113; 1993]. Excellent textbooks have been written on the semismooth Newton method; let us cite [83, 50, 74, 76; 2002-2014].

Exercises

4.4.1. *Strong semismoothness of the ℓ_p -norm.* For $p \in [1, \infty]$, the ℓ_p -norm $\|\cdot\|_p : \mathbb{R}^n \rightarrow \mathbb{R}$, defined at $x \in \mathbb{R}^n$ by

$$\|x\|_p = \begin{cases} (\sum_{1 \leq i \leq n} |x_i|^p)^{1/p} & \text{if } 1 \leq p < \infty, \\ \max_{1 \leq i \leq n} |x_i| & \text{if } p = \infty, \end{cases}$$

is strongly semismooth.

4.4.2. *Strong semismoothness of C-functions.* The C-function “min” (4.43) and of Fischer (4.44) are strongly semismooth.

4.5 Reformulation Methods for Complementarity Problems ▲

Let \mathbb{E} be a Euclidean space, and $F : \mathbb{E} \rightarrow \mathbb{R}^n$ and $G : \mathbb{E} \rightarrow \mathbb{R}^n$ be two smooth functions. The (*nonlinear*) *complementarity problem* (CP) consists in finding a point $x \in \mathbb{E}$ such that

$$0 \leq F(x) \perp G(x) \geq 0. \quad (4.56)$$

This system means that x must be such that $F(x) \geq 0$ and $G(x) \geq 0$, componentwise, and $F(x)^\top G(x) = 0$. A more general setting is presented in (4.56); here, we limit our presentation to the case where the functions F and G take their values in \mathbb{R}^n and the cone K of \mathbb{R}^n is its nonnegative orthant \mathbb{R}_+^n (recall that \mathbb{R}_+^n is self-dual, meaning that its positive dual cone $(\mathbb{R}_+^n)^+$ is \mathbb{R}_+^n). Less or more recent states of the art on the analysis of complementarity problems and numerical methods to solve them can be found in [103, 73, 107, 50, 39, 40, 76].

Occasionally, we shall make reference to the *linear complementarity problem* (LCP) in its standard form, which reads

$$0 \leq x \perp (Mx + q) \geq 0, \quad (4.57)$$

where its unknown is $x \in \mathbb{R}^n$, while $q \in \mathbb{R}^n$ and $M \in \mathbb{R}^{n \times n}$ are its data. It corresponds to the nonlinear complementarity problem (4.56) with $F : x \mapsto Mx + q$ is affine and $G : x \mapsto x$ is the identity operator.

A major difficulty of problem (4.56) (and (4.57)) comes from its combinatorial aspect. Since both $F(x)$ and $G(x)$ must have nonnegative components, the orthogonality conditions $F(x)^\top G(x) = 0$ is equivalent to the n identities

$$\forall i \in [1:n] : \quad F_i(x)G_i(x) = 0. \quad (4.58)$$

There are 2^n possibilities to realize (4.58), by forcing, for each $i \in [1:n]$, either $F_i(x)$ or $G_i(x)$ to vanish. This fact yields much difficulty to the algorithms to find a solution. For instance, even when the functions F and G are affine like in (4.57), finding a solution is NP-hard [31, 84; 1989-1991].

Since at a solution \bar{x} to (4.56), either $F(\bar{x}) = G(\bar{x}) = 0$, or $0 = F(\bar{x}) < G(\bar{x})$, or $F(\bar{x}) > G(\bar{x}) = 0$, it is natural to introduce the following index sets:

$$\begin{aligned} \mathcal{E}(x) &:= \{i \in [1:n] : F_i(x) = G_i(x)\}, \\ \mathcal{F}(x) &:= \{i \in [1:n] : F_i(x) < G_i(x)\}, \\ \mathcal{G}(x) &:= \{i \in [1:n] : F_i(x) > G_i(x)\}. \end{aligned} \quad (4.59)$$

Obviously, these form a partition of $[1:n]$.

The decomposition of the orthogonality condition $F(x)^\top G(x) = 0$ into the n complementarity conditions (4.58) also shows that these count for n equations. Indeed, if the index sets $\mathcal{E}(\bar{x})$, $\mathcal{F}(\bar{x})$, and $\mathcal{G}(\bar{x})$ at a solution \bar{x} of (4.56) were known, this solution would also satisfy the following system of n equations

$$\begin{cases} F_i(x) = 0 & \text{if } i \in \tilde{\mathcal{F}}(\bar{x}), \\ G_i(x) = 0 & \text{if } i \in \tilde{\mathcal{G}}(\bar{x}), \end{cases}$$

where the pair $(\tilde{\mathcal{F}}(\bar{x}), \tilde{\mathcal{G}}(\bar{x}))$ forms a partition of $[1:n]$ and satisfies $\tilde{\mathcal{F}}(\bar{x}) \supseteq \mathcal{F}(\bar{x})$ and $\tilde{\mathcal{G}}(\bar{x}) \supseteq \mathcal{G}(\bar{x})$, together with the implicit constraints $F(x) \geq 0$ and $G(x) \geq 0$. Therefore, the system (4.56) has more chance to be well-posed if $\dim \mathbb{E} = n$. It is often the case that the complementarity system (4.56) is completed by equality constraints; the system is then called a *mixed complementarity problem*; if there is m such equalities, it is natural to have $\dim \mathbb{E} = n + m$.

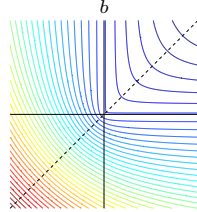
Complementarity conditions arise spontaneously in the first order optimality conditions of an optimization problem with inequality constraints and these conditions can be written like the system (4.56); see (1.58c) in the KKT system or, more generally, (??). The complementarity system (4.56) is also often used to model in part problems in which several systems of equations are, to some extent, in competition. The one that is active in a given place and at a given time, corresponding to a common index of $F(x)$ and $G(x)$, depends on threshold effects; if the threshold $F_i(x) = 0$ is not reached, i.e., $F_i(x) > 0$, then the equation $G_i(x) = 0$ is active, and vice versa. Examples include problems in nonsmooth mechanics and dynamics [1, 25], the phase transition problem in multiphase flows [98, 99, 12, 41, 9, 27, 10, 11], precipitation-dissolution problems in chemistry [26, 85], portfolio management in finance [58], computer graphics [48], meteorology simulation, economic equilibrium, to mention a few. Surveys on examples of applications of the complementarity problem can be found in [63, 73, 107, 52, 50].

Many techniques have been proposed to solve (4.56) since the problem was introduced by Cottle in his PhD thesis, dated 1964 [36, 37]. It is out of the scope of this section to review all of them and we refer instead the interested reader to the monographs [50, 76]. Below, we limit our account to the algorithms using the two most often encountered reformulation of (4.56) in the form of a nonsmooth equation. The reformulation by the Fischer function is examined in section 4.5.1 and the reformulation using the minimum function is considered in sections 4.5.2 and 4.5.2.

4.5.1 Fischer-Newton Algorithm

The Fischer C-Function

In this section, we consider the reformulation based on the Fischer [54] C-function $\varphi_F : \mathbb{R}^2 \rightarrow \mathbb{R}$, defined at $(a, b) \in \mathbb{R}^2$ by

$$\varphi_F(a, b) = \sqrt{a^2 + b^2} - (a + b). \quad (4.60)$$


This is indeed a C-function since, for two real numbers a and b , $\varphi_F(a, b) = 0$ if and only if $a \geq 0$, $b \geq 0$, and $ab = 0$. This function is convex (it is the Euclidean norm plus a linear function) and is C^∞ , except at $(a, b) = (0, 0)$ where it is non differentiable. The function is strongly semismooth however (exercise 4.4.2).

The equation reformulation of (4.56) using φ_F reads

$$\Phi_F(x) = 0, \quad (4.61a)$$

where $\Phi_F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is defined at x by

$$\Phi_F(x) = \varphi_F(F(x), G(x)), \quad (4.61b)$$

where φ_F acts componentwise. A natural merit function associated with the reformulation (4.61) is the least-square function $\theta_F : \mathbb{R}^n \rightarrow \mathbb{R}$, defined at $x \in \mathbb{R}^n$ by

$$\theta_F(x) = \frac{1}{2} \|\Phi_F(x)\|^2,$$

where $\|\cdot\|$ denotes the Euclidean norm.

We have the following smoothness result.

Proposition 4.32 (smoothness of the Fischer reformulation) *If F and G are $C^{1,1}$, then Φ_F is locally Lipschitz and θ_F is $C^{1,1}$.*

Proof. See [51, 30]. □

Globalization

A function $F : \mathbb{E} \rightarrow \mathbb{E}$ is called a *uniform P-function* [29] if there is an $\alpha > 0$ such that for all $x, y \in \mathbb{E}$:

$$\max_{i \in [1 : n]} (y_i - x_i)(F_i(y) - F_i(x)) \geq \alpha \|y - x\|^2.$$

Proposition 4.33 (smoothness of the Fischer reformulation) *Suppose $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a C^1 uniform \mathbf{P} -function. Then, for any $x \in \mathbb{R}^n$ and any $J \in \partial\Phi_{\mathbf{F}}(x)$, J is nonsingular.*

Proof. See [77; 1997, proposition 3.2]. □

4.5.2 Newton-Min Algorithm

The Minimum C-Function

In this section, we consider the reformulation based on the Minimum C-function $\varphi_{\min} : \mathbb{R}^2 \rightarrow \mathbb{R}$, defined at $(a, b) \in \mathbb{R}^2$ by

$$\varphi_{\min}(a, b) = \min(a, b). \tag{4.62}$$

This is indeed a C-function since, for two real numbers a and b , $\varphi_{\min}(a, b) = 0$ if and only if $a \geq 0$, $b \geq 0$, and $ab = 0$. This function is concave (minimum of two linear functions) and strongly semismooth (exercise 4.4.2).

Plain Newton-Min algorithm

Polyhedral Newton-Min Algorithm

Notes

There are many other *C-functions* than those used in sections 4.5.1 and 4.5.2 that have been proposed in the literature. Tseng [133] studies the C-function obtained by replacing the ℓ_2 norm of (a, b) in the Fischer Function by its ℓ_p norm.

5 A Few Methods for Optimization

5.1 SQP Algorithm for (P_{EI})

In this section, we apply the Josephy-Newton (JN) algorithm of section 4.1.2 to the equality and inequality constrained optimization problem (P_{EI}) , more precisely to the system formed by its first order optimality conditions. As already observed in section 4.1.1, this optimality system can indeed be written as a complementarity problem, which is a special case of function inclusion. This approach yields an algorithm named SQP for *Sequential Quadratic Programming*, which is one of the most often used algorithm to solve problem (P_{EI}) , when derivatives are available.

In addition to give a consistent and illuminating way of introducing the SQP algorithm, this approach is also fruitful. In particular, it offers to possibility to get the conditions of local convergence inherited from those ensuring the local convergence of the JN algorithm (theorem 4.12), which were not known before this technique was introduced in [19; 1994].

For the reader's convenience, we recall the form of the optimization problem (P_{EI}) :

$$(P_{EI}) \quad \begin{cases} \min f(x) \\ c_i(x) = 0, & i \in E \\ c_i(x) \leq 0, & i \in I. \end{cases}$$

In this setting, the vector x to optimize lies in a Euclidean vector space \mathbb{E} and the functions $f : \mathbb{E} \rightarrow \mathbb{R}$ and $c_i : \mathbb{E} \rightarrow \mathbb{R}$ defining the objective and the constraints are supposed smooth. The index sets E and I form a partition of $[1 : m]$: $E \cup I = [1 : m]$ and $E \cap I = \emptyset$.

Here are some more notation. We denote by $c : \mathbb{E} \rightarrow \mathbb{R}^m$ the function whose i th component is c_i . The cardinality of E and I are denoted by $m_E := |E|$ and $m_I := |I|$, so that $m = m_E + m_I$. If $v \in \mathbb{R}^m$, we denote by v_E (resp. v_I) the vector of \mathbb{R}^{m_E} (resp. \mathbb{R}^{m_I}) formed of the components v_i of v with index $i \in E$ (resp. $i \in I$). Applying this to c , $c_E : \mathbb{E} \rightarrow \mathbb{R}^{m_E}$ (resp. $c_I : \mathbb{E} \rightarrow \mathbb{R}^{m_I}$) is now the constraint function defining the equality (resp. inequality) constraints. To a vector $v \in \mathbb{R}^m$, we associate the vector $v^\# \in \mathbb{R}^m$, defined by

$$(v^\#)_i = \begin{cases} v_i & \text{if } i \in E \\ v_i^+ & \text{if } i \in I, \end{cases}$$

where $v_i^+ = \max(0, v_i)$. With this notation, the constraints of (P_{EI}) read $c(x)^\# = 0$, whose interest lies in its compactness (note indeed that $x \mapsto c(x)^\#$ is usually nonsmooth, so that the difficulty associated with the inequality constraints has been transferred to the difficulty coming from nonsmoothness).

5.1.1 The SQP Algorithm

Figure 5.1 below gives a flowchart that can allow the reader to see schematically the

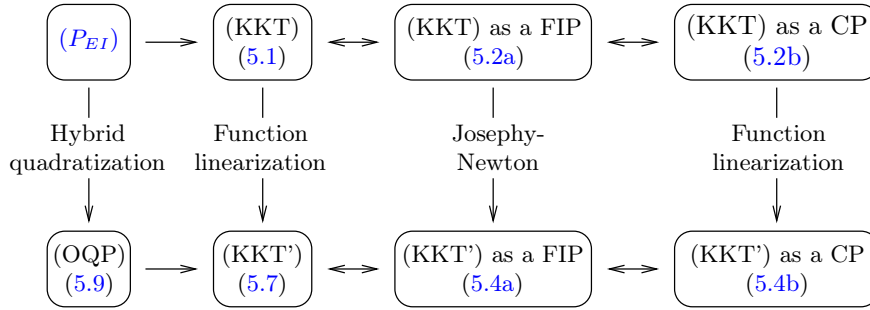


Fig. 5.1. Flowchart of the systems encountered in section 5.1.1.

links between all the systems encountered in this section, which lead to the definition of the SQP algorithm and its *osculating quadratic problem* (OQP). The meaning of its blocs, which refer to formulas not already encountered, will be revealed progressively throughout the section. A simple horizontal arrow indicates an implication between the systems; a double horizontal arrow means an equivalence. The labels of the vertical arrows indicate the type of transformation used to go from the upper box to the lower one. Even though we do not have all the elements at hand to understand this diagram, we can already outline it.

1. Most presentations of the SQP algorithm confine themselves to the leftmost two blocs, hence viewing the osculating quadratic problem (OQP) defined below as a kind of hybrid quadratization of (P_{EI}) .
2. When, one goes through the first order optimality conditions, the (KKT) bloc in the flowchart, one gets the first order optimality conditions (KKT') of the OQP by a pseudo-linearization, which linearizes the functions in the KKT system, while keeping its structure. This is probably the fastest and meaningful way of presenting the SQP algorithm and its osculating quadratic problem.
3. It makes even more sense, and this is what we do below, to view the KKT system as the functional inclusion problem (5.2a), denoted FIP in the flowchart, and apply the JN linearization to it. Then, one gets (5.4a), which is actually a functional inclusion expression of (KKT').
4. The approach made in point 3 on the functional inclusion form (5.2a) of (KKT), can also be done on its complementarity form (5.2b). The linearization of the functions appearing in the latter yields (5.4b), which is the complementarity form of (KKT').

We now expound the approach described schematically in the points 3 and 4 aforementioned, with more precision.

Like Newton's method for solving an unconstrained optimization method, the SQP algorithm for solving (P_{EI}) focuses on the solutions to the first order optimality

conditions of the problem or its KKT system (1.58), which we recall below for the reader's convenience. It is the following system in $(x, \lambda) \in \mathbb{E} \times \mathbb{R}^m$:

$$\text{(KKT)} \quad \begin{cases} \nabla f(x) + c'(x)^* \lambda = 0 \\ c_E(x) = 0 \\ 0 \leq \lambda_I \perp -c_I(x) \geq 0. \end{cases} \quad (5.1)$$

With the inequality $-c_I(x) \geq 0$, the last complementarity condition is written slightly differently than in (1.58), since for the sequel we would like to have λ_I and $-c_I(x)$ to belong to dual cones (actually to $\mathbb{R}_+^{m_I}$, which is **self-dual** in the sense that $(\mathbb{R}_+^{m_I})^+ = \mathbb{R}_+^{m_I}$). A pair (x, λ) solution to this system is called a *primal-dual solution* and the primal solution x is also sometimes called a *stationary point*.

We have seen in section 4.1.1, and this is a fundamental observation, that this system reads like the following functional inclusion or complementarity problem in $z := (x, \lambda)$:

$$F(z) + N_{K^+}(z) \ni 0, \quad (5.2a)$$

$$K^+ \ni z \perp F(z) \in K, \quad (5.2b)$$

where

$$F(z) = \begin{pmatrix} \nabla f(x) + c'(x)^* \lambda \\ -c(x) \end{pmatrix} \quad \text{and} \quad K = \{0_{\mathbb{E}}\} \times (\{0_{\mathbb{R}^{m_E}}\} \times \mathbb{R}_+^{m_I}). \quad (5.3)$$

Because of the importance of this equivalence for this section, let us check it again. Since $K^+ = \mathbb{E} \times (\mathbb{R}^{m_E} \times \mathbb{R}_+^{m_I})$, we have that

- $(x, \lambda) \in K^+$ reads $\lambda_I \geq 0$,
- $F(x, \lambda) \in K$ reads $\nabla_x \ell(x, \lambda) = 0$, $c_E(x) = 0$, and $c_I(x) \leq 0$,
- $(x, \lambda) \perp F(x, \lambda)$ amounts then to the complementarity expression $\lambda_I \perp c_I(x)$.

Hence, one recovers indeed (5.1).

The pure **JN** algorithm (4.11), that is with $M_k = F'(x_k)$, applied to the functional inclusion (5.2a) or to the complementarity problem in (5.2b) leads to determine the next iterate $z_{k+1} := (x_{k+1}, \lambda_{k+1})$ from the current one $z_k := (x_k, \lambda_k)$ by solving in $z := (x, \lambda)$ the following linearized functional inclusion problem or its equivalent linearized complementarity problem:

$$F(z_k) + F'(z_k)(z - z_k) + N_{K^+}(z) \ni 0, \quad (5.4a)$$

$$K^+ \ni z \perp (F(z_k) + F'(z_k)(z - z_k)) \in K. \quad (5.4b)$$

Let us see the form of this algorithm when (F, K) is given by (5.3). Observe that

$$F'(z) = \begin{pmatrix} L(x, \lambda) & c'(x)^* \\ -c'(x) & 0 \end{pmatrix}, \quad (5.5)$$

where we simplified the notation by introducing

$$L(x, \lambda) := \nabla_{xx}^2 \ell(x, \lambda).$$

Consider (5.4b).

- The condition $z \in K^+$ reads like above

$$\lambda_I \geq 0. \quad (5.6a)$$

- The condition $F(z_k) + F'(z_k)(z - z_k) \in K$ means

$$\nabla_x \ell(x_k, \lambda) + L(x_k, \lambda_k)(x - x_k) + c'(x_k)^*(\lambda - \lambda_k) = 0, \quad (5.6b)$$

$$c_E(x_k) + c'_E(x_k)(x - x_k) = 0, \quad (5.6c)$$

$$c_I(x_k) + c'_I(x_k)(x - x_k) \leq 0. \quad (5.6d)$$

- Finally, the orthogonality relation $z \perp (F(z_k) + F'(z_k)(z - z_k))$ can be expressed by

$$\lambda_I \perp [c_I(x_k) + c'_I(x_k)(x - x_k)]. \quad (5.6e)$$

Gathering the conditions in (5.6) yields the system

$$(KKT') \quad \begin{cases} \nabla_x \ell(x_k, \lambda_k) + L(x_k, \lambda_k)(x - x_k) + c'(x_k)^*(\lambda - \lambda_k) = 0, \\ c_E(x_k) + c'_E(x_k)(x - x_k) = 0, \\ 0 \leq \lambda_I \perp [c_I(x_k) + c'_I(x_k)(x - x_k)] \leq 0. \end{cases} \quad (5.7)$$

We could obtain the same system from the KKT system (5.1) by linearizing its functions and keeping its structure, made of the equalities, inequalities and perpendicularity operator.

Solving (KKT') is not an easy task: this system is not a classic problem that can be solved by the usual pieces of software; in addition, the original optimization nature of problem (P_{EI}) looks lost in this formulation. This latter observation is in appearance only, since the fact that the system (KKT') comes from the linearization of a KKT system, it is also a KKT system, but now of a quadratic optimization problem. Let us see this. Note first that it is not the I -component of the multiplier $(\lambda - \lambda_k)$ appearing in the first equation of (5.7) that must be nonnegative, but λ_I , as imposed by the last condition in (5.7). This is problematic when one tries to see (5.7) as a KKT system. But there is cure to that difficulty, which consists in eliminating λ_k from the first equation of (5.7) (it appears twice in terms that cancel each other). The system becomes

$$\begin{cases} \nabla f(x_k) + L(x_k, \lambda_k)(x - x_k) + c'(x_k)^*\lambda = 0, \\ c_E(x_k) + c'_E(x_k)(x - x_k) = 0, \\ 0 \leq \lambda_I \perp [c_I(x_k) + c'_I(x_k)(x - x_k)] \leq 0. \end{cases} \quad (5.8)$$

Now, it is not difficult to observe that (5.8) is formed of the first order optimality conditions of the following quadratic optimization problem

$$\begin{cases} \min_x \langle \nabla f(x_k), (x - x_k) \rangle + \frac{1}{2} \langle L(x_k, \lambda_k)(x - x_k), (x - x_k) \rangle \\ c_E(x_k) + c'_E(x_k)(x - x_k) = 0 \\ c_I(x_k) + c'_I(x_k)(x - x_k) \leq 0. \end{cases}$$

Note that the multiplier associated with the linearized constraints is the sought dual solution λ to this problem, not the increment $\lambda - \lambda_k$. For the sequel, it is convenient to set $d = x - x_k$, so that the previous quadratic problem can also be written as a quadratic problem in $d \in \mathbb{E}$:

$$(\text{OQP})_k \quad \begin{cases} \min_d \langle \nabla f(x_k), d \rangle + \frac{1}{2} \langle \nabla_{xx}^2 \ell(x_k, \lambda_k) d, d \rangle \\ c_E(x_k) + c'_E(x_k) d = 0 \\ c_I(x_k) + c'_I(x_k) d \leq 0. \end{cases} \quad (5.9)$$

This problem is called the *osculating quadratic problem* of (P_{EI}) at (x_k, λ_k) . Its objective has a hybrid nature, since the linear term is formed with the gradient of the objective of (P_{EI}) , while the quadratic term is formed with the Hessian of the Lagrangian of the problem.

One can now specify the *local SQP* iteration, local means here *without globalization technique* like those of sections 5.1.4 and 5.1.5.

Algorithm 5.1 (local SQP) One iteration, from $(x_k, \lambda_k) \in \mathbb{E} \times \mathbb{R}^m$ to (x_{k+1}, λ_{k+1}) is made of the following steps:

1. *Stopping test*: if the current pair (x_k, λ_k) is satisfactory, stop;
2. *QP solve*: let (d_k, λ_{k+1}) be an *appropriate* primal-dual solution to the osculating quadratic problem (5.9), if any;
3. *New iterate*: set $x_{k+1} := x_k + d_k$ and $\lambda_{k+1} := \lambda_{k+1}$.

This algorithm deserves some remarks.

Remarks 5.2 1) The SQP algorithm decomposes the computation of a solution to (P_{EI}) in a sequence of osculating quadratic optimization problems (5.9), easier to solve than (P_{EI}) . Therefore, the combinatorial aspect of problem (P_{EI}) (linked to the determination of the active inequality constraints), is transferred to the OQP, where it is still serious, but less than in the nonlinear problem (P_{EI}) .

2) The computationally expensive part of the SQP algorithm is the computation of the solution to the osculation quadratic problem (OQS), which can be much more computation time consuming than a linear system. Therefore, we are in the class of algorithms with an expensive iteration; this is expected since we have seen that the SQP algorithm is derived from the JN algorithm, which has that property.

3) The OQP is still computationally expensive to solve:

- if $L_k \not\geq 0$, then the OQP is NP-hard,
- if $L_k \geq 0$, then the OQP can be solved in polynomial time (by an interior point method, but this is not necessarily the best approach).

For this reason, many implementations approach the Hessian of the Lagrangian L_k by a positive definite matrix (using a quasi-Newton technique for example), see also section 5.1.4.

4) We shall see in section 5.1.2 that algorithm 5.1 enjoys a local rapid convergence: it is quadratic if f and c are sufficiently smooth (of class $\mathcal{C}^{2,1}$). This means that, once an iterate is close to a “regular solution” (a notion that will be clarify in the next section), the convergence to that solution is very fast (less than 5 or 10 iterations, to give a number), whatever the dimension of the problem is. Furthermore, the algorithm provides a very accurate approximation to the solution. Nothing is done, however, in algorithm 5.1 to ensure the convergence if the itnitial iterate (x_1, λ_1) is not close to a regular solution. This subject is considered in sections 5.1.4 and 5.1.5.

- 5) The phrase “if any” in step 2 of the algorithm hides a lot of difficulties that a fully developed piece of software must overcome. Let us mention the three main ones.
- It may occur that the linearized constraints of the OQP are not compatible, even if problem (P_{EI}) is feasible. For example, assume that in (P_{EI}) , $n = 1$, $m_E = 0$, and $m_I = 2$, with the constraints

$$x \geq 0 \quad \text{and} \quad \log(x + 1) \leq 1,$$

which is a complicated way of requiring to have $x \in [0, 9]$. The linearization of these compatible constraints at $x = 99$ reads “ $x \geq 0$ and $2 + (x - 99)/100 \leq 1$ ” or “ $x \geq 0$ and $x \leq -1$ ”, which are not compatible.

- It may occur that the OQP is feasible but unbounded (its optimal value is $-\infty$).
- If none of the previous situations occur, the OQS has a solution (like in linear optimization, since one can show that *a possibly nonconvex quadratic optimization problem with a real optimal value has a solution* [56; 1956, appendix (i)]). Nevertheless, it may have undesirable solutions. Here is an example. Consider the following problem in $x \in \mathbb{R}$:

$$\begin{cases} \min_x \log(x + 1) \\ 0 \leq x \leq 3. \end{cases}$$

Its solution is $x_* = 0$ and there is a unique associated multiplier $\lambda_* = (1, 0)$. The OQP at (x_*, λ_*) reads

$$\begin{cases} \min_d d - \frac{1}{2} d^2 \\ 0 \leq d \leq 3. \end{cases}$$

This problem has three stationary points: 0, which is a local minimum, 1, which is a global maximum, and 3, which is a global minimum. Clearly, only the first one is satisfactory (at a solution to a problem any sensible algorithm should provide a zero displacement).

5.1.2 Local Convergence

Definition 5.3 A stationary pair $z_* := (x_*, \lambda_*)$ of (P_{EI}) is said to be *semi-stable* (resp. *hemi-stable*) if z_* is a semi-stable (resp. hemi-stable) solution to the functional inclusion (5.2a) with F and K given by (5.3).

If (x_*, λ_*) verifies the KKT conditions, it follows that

$$F(x_*, \lambda_*) = \begin{pmatrix} 0_E \\ 0_{\mathbb{R}^{m_E}} \\ -c_I(x_*) \end{pmatrix}, \tag{5.10}$$

$$F(x_*, \lambda_*) + F'(x_*, \lambda_*)(d, \mu) = \begin{pmatrix} L_* d + c'(x_*)^* \mu \\ -c'_E(x_*) d \\ -c_I(x_*) - c'_I(x_*) d \end{pmatrix}. \tag{5.11}$$

where we have used (5.5) and set $L_* := L(x_*, \lambda_*) := \nabla_{xx}^2 \ell(x_*, \lambda_*)$. The condition $F(x_*, \lambda_*) + F'(x_*, \lambda_*)(d, \mu) + N_{K^+}(x_*, \lambda_*) \ni 0$ reads

$$L_*d + c'(x_*)^*\mu = 0, \\ c'_{E \cup I_*^{0+}}(x_*)d = 0, \quad c'_{I_*^{00}}(x_*)d \leq 0, \quad c_{I_*^\sim}(x_*) + c'_{I_*^\sim}(x_*)d \leq 0.$$

Proposition 5.4 (semi-stability of a local minimum) *If x_* is a local minimum of (P_{EI}) and λ_* is an associated optimal multiplier, then the following properties are equivalent:*

- (i) (x_*, λ_*) is semi-stable,
- (ii) λ_* is the unique optimal multiplier associated with x_* and the second order sufficient conditions of optimality hold: $\langle L_*d, d \rangle > 0$, for all $d \in C_* \setminus \{0\}$.

Proof. 1) We take advantage of the equivalence (i) \Leftrightarrow (iii) of proposition 4.7 to get another expression of the semi-stability of (x_*, λ_*) , hence another expression of point (i) in the present proposition. To this end, we introduce

$$(d, \mu) := z - z_* = (x - x_*, \lambda - \lambda_*)$$

and we observe that, with F defined in (5.3), with (5.10) and (5.11), one has

$$\langle F'(x_*, \lambda_*)(d, \mu), (d, \mu) \rangle = \langle L_*d, d \rangle, \\ \langle F(x_*, \lambda_*), (d, \mu) \rangle = 0 \iff \mu_{I_*^\sim} = 0.$$

By the equivalence (i) \Leftrightarrow (iii) of proposition 4.7, the semi-stability of (x_*, λ_*) is equivalent to

$$\text{one has } \langle L_*d, d \rangle > 0 \text{ for all } (d, \mu) \in \mathbb{E} \times \mathbb{R}^m \text{ such that} \quad (5.12a)$$

$$(d, \mu) \neq 0, \quad (5.12b)$$

$$(\lambda_* + \mu)_{I_*^0} \geq 0, \quad \mu_{I_*^\sim} = 0, \quad (5.12c)$$

$$L_*d + c'_{E \cup I_*^0}(x_*)^*\mu_{E \cup I_*^0} = 0, \quad (5.12d)$$

$$d \in C_*, \quad \text{and} \quad c_{I_*^\sim}(x_*) + c'_{I_*^\sim}(x_*)d \leq 0. \quad (5.12e)$$

where $C_* := \{d \in \mathbb{E} : c'_{E \cup I_*^0}(x_*)d = 0, c'_{I_*^0}(x_*)d \leq 0\}$ is the critical cone.

2) [(i) \Rightarrow (ii)] Since (x_*, λ_*) is semi-stable, it is an *isolated* solution to the optimality system (1.58) (proposition 4.4). Since the set of optimal multipliers associated with x_* is convex (it is a convex polyhedron), this one must be a singleton, which proves the first part of (ii).

The set of optimal multipliers associated with x_* being bounded (it is a singleton), the constraint qualification condition (CQ-MF) holds (see proposition 2.22 and exercise 2.1.4). Therefore, by the second order necessary conditions of optimality (theorem 2.37), $\langle L_*d, d \rangle \geq 0$ for all critical directions $d \in C_*$. To show that the second order sufficient conditions of optimality also holds, we proceed by contradiction, assuming that there is a nonzero direction $d_1 \in C_*$ such that $\langle L_*d_1, d_1 \rangle = 0$. Then, this direction d_1 is a solution to the quadratic problem

$$\begin{cases} \min \langle L_*d, d \rangle \\ c'_{E \cup I_*^0}(x_*)d = 0 \\ c'_{I_*^0}(x_*)d \leq 0, \end{cases}$$

whose constraints define critical directions. Since (CQ-A) holds for this problem, its optimality conditions ensure the existence of a multiplier $\mu_1 \in \mathbb{R}^m$ such that

$$\begin{aligned} (\mu_1)_{I_*^c} &= 0, \\ L_* d_1 + c'(x_*)^* \mu_1 &= 0, \\ c'_{E \cup I_*^{0+}}(x_*) d_1 &= 0, \\ 0 &\leq (\mu_1)_{I_*^{00}} \perp c'_{I_*^{00}}(x_*) d_1 \leq 0. \end{aligned}$$

Then $(d, \mu) = t(d_1, \mu_1)$, for $t > 0$ sufficiently small, verifies (5.12b)-(5.12e) but not the conclusion $\langle L_* d, d \rangle > 0$ in (5.12a). This contradiction shows that SC2 is verified.

3) [(i) \Leftarrow (ii)] To show the semi-stability of (x_*, λ_*) , we show that (5.12) holds. Let (d, μ) verifying (5.12b)-(5.12e). It suffices to show that $d \neq 0$, since then $d \in C_* \setminus \{0\}$ and the conclusion $\langle L_* d, d \rangle > 0$ in (5.12a) follows from the SC2 that is assumed in (ii). We proceed again by contradiction, assuming that $d = 0$. Then $\mu \neq 0$ by (5.12b) and it is plain to see by (5.12c)-(5.12d) that $\lambda_* + \mu$ would then be another optimal multiplier associated with x_* , which would contradict the uniqueness of the optimal multiplier, assumed in (ii). \square

Proposition 5.5 (sufficient condition of hemi-stability) *If x_* is a local minimum of (PEI) , with an associate multiplier λ_* such that (x_*, λ_*) is semi-stable, then (x_*, λ_*) is also hemi-stable.*

Proof. By the definition 4.11 of the hemi-stability, one has to show that, for all $\alpha > 0$, one can find $\beta > 0$, such that, for all $(x_0, \lambda_0) \in \bar{B}((x_*, \lambda_*), \beta)$, the following inclusion in (x, λ)

$$\begin{pmatrix} \nabla f(x_0) + c'(x_0)^* \lambda_0 \\ -c(x_0) \end{pmatrix} + \begin{pmatrix} L(x_0, \lambda_0) & c'(x_0)^* \\ -c'(x_0) & 0 \end{pmatrix} \begin{pmatrix} x - x_0 \\ \lambda - \lambda_0 \end{pmatrix} + N_{K^+}(x, \lambda) \ni 0$$

has a solution in $\bar{B}((x_*, \lambda_*), \alpha)$. This inclusion is the first order optimality system of the following quadratic optimization problem in $x \in \mathbb{E}$:

$$\begin{cases} \min \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2} \langle L(x_0, \lambda_0)(x - x_0), x - x_0 \rangle \\ c_E(x_0) + c'_E(x_0)(x - x_0) = 0 \\ c_I(x_0) + c'_I(x_0)(x - x_0) \leq 0. \end{cases} \quad (5.13)$$

The system (5.13) can be viewed as a perturbation of the quadratic optimization problem in $x \in \mathbb{E}$ that is obtained by taking $(x_0, \lambda_0) = (x_*, \lambda_*)$, namely

$$\begin{cases} \min \langle \nabla f(x_*), x - x_* \rangle + \frac{1}{2} \langle L_*(x - x_*), x - x_* \rangle \\ c_E(x_*) + c'_E(x_*)(x - x_*) = 0 \\ c_I(x_*) + c'_I(x_*)(x - x_*) \leq 0. \end{cases} \quad (5.14)$$

The first order optimality conditions of (5.14) read: there exists $\lambda \in \mathbb{R}^m$ such that

$$\begin{aligned}\nabla f(x_*) + L_*(x - x_*) + c'(x_*)^* \lambda &= 0 \\ c_E(x_*) + c'_E(x_*)(x - x_*) &= 0 \\ 0 \leq \lambda \perp (c_I(x_*) + c'_I(x_*)(x - x_*)) &\leq 0.\end{aligned}$$

By the KKT conditions of problem (P_{EI}) , this system is verified by $(x, \lambda) = (x_*, \lambda_*)$. Now, by proposition 5.4, the assumed semi-stability of (x_*, λ_*) implies that λ_* is the unique optimal multiplier associated with x_* and that the second order sufficient conditions of optimality hold. We deduce from this that first (x_*, λ_*) verifies the second order optimality conditions of (5.14), which are the same as those of (P_{EI}) , and that (CQ-MF) holds for (5.14) (by the uniqueness of the associated multiplier and the Gauvin property of proposition 2.22). By proposition 3.1, these properties ensure that the perturbed problem (5.13) has a primal-dual solution (x, λ) (it may have other undesirable solutions, however), whose distance to (x_*, λ_*) is bounded by a constant times the norm of the perturbation $(x_0 - x_*, \lambda_0 - \lambda_*)$. \square

Theorem 5.6 (local convergence of the SQP algorithm) *If f and c are $\mathcal{C}^{2,1}$ in a neighborhood of a local minimum x_* of (P_{EI}) , if there exists a unique multiplier associated with x_* , and if the sufficient conditions of optimality of the second order are satisfied, then there exists a neighborhood V of (x_*, λ_*) such that, if the first iterate $(x_1, \lambda_1) \in V$, then*

- 1) *the SQP algorithm can generate a sequence $\{(x_k, \lambda_k)\}$ in V ,*
- 2) *$\{(x_k, \lambda_k)\}$ converges quadratically to (x_*, λ_*) .*

Proof. By proposition 5.4, the uniqueness of the optimal multiplier, and the second order optimality conditions, (x_*, λ_*) is a semi-stable solution of (5.2). By proposition 5.5, this is also an hemi-stable solution. One can then apply theorem 4.12, which gives the result. \square

5.1.3 Exact Penalization

Motivation

A review of penalty techniques

The augmented Lagrangian is a first way of getting an exact penalization, provided one knows an optimal multiplier (since this is usually not the case, the multiplier method generates a sequence approaching an optimal multiplier). The underlying idea is to penalize quadratically a function whose first derivative vanishes at the considered solution, which is the Lagrangian $\ell(\cdot, \lambda_*)$.

Another way of getting an exact penalty function is to do this using a nondifferentiable function. Let us illustrate the idea in the case of the following simple optimization problem in $x \in \mathbb{R}$:

$$\begin{cases} \inf 1 - x - \frac{1}{3}x^3 \\ x \leq 0. \end{cases} \quad (5.15)$$

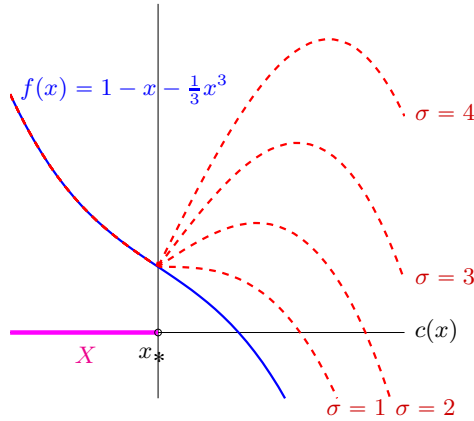


Fig. 5.2. Nondifferentiable penalization for the problem (5.15) with $\sigma = 1, 2, 3$ and 4.

Figure 5.2 shows that if the term $\sigma x^+ = \sigma \max(0, x)$ is added to the objective of the problem, one gets an exact penalty function, as soon as $\sigma > 1$. This threshold $\sigma = 1$ comes in the present case from the slope of f at zero. More generally, it is the “slope” (if this one exists) of the value function at zero that is important, so that it is the optimal multipliers associated with x_* that will play a key role in the determination of the value of the threshold above which an exact penalty is obtained.

An exact penalty function

In this section, we consider the following penalty function, which is associated with problem (P_{EI}) :

$$\Theta_\sigma(x) = f(x) + \sigma \|c(x)\|_P, \tag{5.16}$$

where $\sigma > 0$ and $\|\cdot\|_P$ is an arbitrary norm. The fundamental result is given in theorem 5.11; it provides conditions ensuring the exactness of Θ_σ . Its proof uses the following three lemmas.

The first lemma is relevant in a larger context than ours, since it highlights conditions for having the directional differentiability of a composition of functions and shows that the chain rule (5.17) applies in that case. To motivate the Lipschitz assumption taken in the lemma, let us point out that the composition of functions having directional derivatives may not have a directional derivative.

Counter-example 5.7 (not directionally differentiable composition) Let $\varphi : \mathbb{R} \rightarrow \mathbb{R}^2$ and $\psi : \mathbb{R}^2 \rightarrow \mathbb{R}$ be defined by

$$\varphi(x) = \begin{cases} (x, x^2 \sin(1/x)) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases} \quad \text{and} \quad \psi(y_1, y_2) = \begin{cases} y_1 & \text{if } y_2 = 0 \\ 0 & \text{if } y_2 \neq 0. \end{cases}$$

It is plain to see that φ is Fréchet differentiable at zero, that ψ is positively homogeneous, hence directionally differentiable at zero, but that $\psi \circ \varphi$ is not directionally differentiable at zero. □

This counter-example [127; p. 484] shows that the Lipschitz continuity of the second function of the composition, the function ψ , assumed in the lemma below, is not superfluous. It is worth noting that the notion of *directional differentiability in the sense of Hadamard* (i.e., $[f(x+t_k d_k) - f(x)]/t_k$ converges to the same vector whatever the sequences $\{d_k\} \rightarrow d$ and $\{t_k\} \downarrow 0$ are) is stable for the composition and the chain rule applies [127; proposition 3.6].

Lemma 5.8 (directional differentiability of a composition) *Let \mathbb{E} , \mathbb{F} , and \mathbb{G} be three normed vector spaces. Suppose that $\varphi : \mathbb{E} \rightarrow \mathbb{F}$ has a directional derivative at $x \in \mathbb{E}$ in the direction $h \in \mathbb{E}$ and that $\psi : \mathbb{F} \rightarrow \mathbb{G}$ is Lipschitz continuous in a neighborhood of $\varphi(x)$ and has a directional derivative at $\varphi(x)$ in the direction $\varphi'(x; h)$. Then $(\psi \circ \varphi)$ has a directional derivative at x in the direction h and there holds*

$$(\psi \circ \varphi)'(x; h) = \psi'(\varphi(x); \varphi'(x; h)). \quad (5.17)$$

Proof. For $t \downarrow 0$, use successively the directional differentiability of φ , the Lipschitz continuity of ψ , and the directional differentiability of ψ :

$$\begin{aligned} (\psi \circ \varphi)(x + th) &= \psi(\varphi(x) + t\varphi'(x; h) + o(t)) \\ &= \psi(\varphi(x) + t\varphi'(x; h)) + o(t) \\ &= (\psi \circ \varphi)(x) + t\psi'(\varphi(x); \varphi'(x; h)) + o(t). \end{aligned}$$

The result follows. \square

The second lemma explores the differentiability of Θ_σ and uses the operator $P_v : \mathbb{R}^m \rightarrow \mathbb{R}^m$, defined for $u \in \mathbb{R}^n$ and $v \in \mathbb{R}_-$ by

$$(P_v u)_i = \begin{cases} u_i & \text{if } i \in E, \\ u_i^+ & \text{if } i \in I \text{ and } v_i = 0, \\ 0 & \text{if } i \in I \text{ and } v_i < 0. \end{cases}$$

to have an explicit expression of $\Theta'_\sigma(x; d)$ at a point x that is feasible or (P_{EI}) .

Lemma 5.9 *If f and c have directional derivatives at $x \in \mathbb{E}$, then Θ_σ has directional derivatives at x . In particular, if x is feasible for (P_{EI}) , the following formula holds*

$$\Theta'_\sigma(x; d) = f'(x; d) + \sigma \|P_{c(x)} c'(x; d)\|_{\mathbb{P}}.$$

Proof. The directional differentiability of $\Theta_\sigma = f + \sigma(\|\cdot\|_{\mathbb{P}} \circ (\cdot)^\# \circ c)$ comes from lemma 5.8, the assumptions on f and c , and the fact that $\|\cdot\|_{\mathbb{P}}$ and $(\cdot)^\#$ are Lipschitz continuous and have directional derivatives.

If x is feasible, $c(x)^\# = 0$ and we have from lemma 5.8,

$$\Theta'_\sigma(x; d) = f'(x; d) + \sigma(\|\cdot\|_{\mathbb{P}})'(0; (c^\#)'(x; d)).$$

We now observe that

$$(\|\cdot\|_{\mathbb{P}})'(0; v) = \lim_{t \rightarrow 0^+} \frac{1}{t} (\|tv\|_{\mathbb{P}} - 0) = \|v\|_{\mathbb{P}}$$

and

$$(c^{\#})'(x; d) = (\cdot^{\#})'(c(x); c'(x; d)) = P_{c(x)} c'(x; d).$$

The result follows. □

The third lemma shows that, when σ is sufficiently large, Θ_{σ} dominates the Lagrangian $\ell(\cdot, \lambda_{*})$ on \mathbb{E} (λ_{*} is an optimal multiplier associated with x_{*}). Recall that the **dual norm** of $\|\cdot\|_{\mathbb{P}}$ for the Euclidean scalar product is the norm $\|\cdot\|_{\mathbb{D}} : \mathbb{R}^m \rightarrow \mathbb{R}$ defined at $v \in \mathbb{R}^m$ by

$$\|v\|_{\mathbb{D}} = \sup_{\|u\|_{\mathbb{P}} \leq 1} v^{\top} u.$$

The *generalized Cauchy-Schwarz inequality*

$$\forall u, v \in \mathbb{R}^m : |u^{\top} v| \leq \|u\|_{\mathbb{P}} \|v\|_{\mathbb{D}}. \tag{5.18}$$

follows readily from the definition of the dual norm.

Lemma 5.10 *If $\lambda \in \mathbb{R}^m$ satisfies $\sigma \geq \|\lambda\|_{\mathbb{D}}$ and $\lambda_I \geq 0$, then, for all $x \in \mathbb{E}$, one has $\ell(x, \lambda) \leq \Theta_{\sigma}(x)$.*

Proof. We have successively

$$\begin{aligned} \ell(x, \lambda) &= f(x) + \lambda^{\top} c(x) && \text{[definition of the Lagrangian]} \\ &\leq f(x) + \lambda^{\top} c(x)^{\#} && [\lambda_I \geq 0 \text{ and } c(x) \leq c(x)^{\#}] \\ &\leq f(x) + \|\lambda\|_{\mathbb{D}} \|c(x)^{\#}\|_{\mathbb{P}} && \text{[(5.18)]} \\ &\leq f(x) + \sigma \|c(x)^{\#}\|_{\mathbb{P}} && [[\|\lambda\|_{\mathbb{D}} \leq \sigma] \\ &= \Theta_{\sigma}(x). \end{aligned} \quad \square$$

Here is the announced result giving sufficient conditions ensuring the exactness of Θ_{σ} at a solution x_{*} of (PEI) . Assumption (5.19) can only hold if the set of optimal multipliers Λ_{*} is bounded, which amounts to say that the Mangasarian-Fromovitz sufficient qualification condition (CQ-MF) holds (see below definition 1.38, exercise 2.1.4 and Gauvin's property of proposition 2.22).

Theorem 5.11 (exactness of Θ_{σ}) *Let x_{*} be a local minimum of (PEI) . Suppose that f and c are twice differentiable at x_{*} and Lipschitz continuous in a neighborhood of x_{*} . Suppose also that x_{*} satisfies the weak second order conditions of optimality (2.80) and denote by Λ_{*} the nonempty set of optimal multi-*

pliers associated with x_* . Suppose finally that

$$\sigma \geq \sup_{\lambda_* \in \Lambda_*} \|\lambda_*\|_{\mathbb{D}} \quad \text{and} \quad \sigma > \|\hat{\lambda}_*\|_{\mathbb{D}}, \text{ for some } \hat{\lambda}_* \in \Lambda_*. \quad (5.19)$$

Then, x_* is a strict local minimum of the penalty function Θ_σ given by (5.16).

Proof. We prove the result by contradiction, assuming that x_* is not a strict local minimum of Θ_σ . Then, there exists a sequence $\{x_k\}$ such that $x_k \neq x_*$, $x_k \rightarrow x_*$ and

$$\Theta_\sigma(x_k) \leq \Theta_\sigma(x_*), \quad \forall k \geq 1. \quad (5.20)$$

We now want to show that the preceding inequalities imply that x_k must approach x_* along a nonzero critical direction.

Since the sequence $\{(x_k - x_*)/\|x_k - x_*\|\}$ is bounded (here $\|\cdot\|$ denotes an arbitrary norm), it has a subsequence such that $(x_k - x_*)/\|x_k - x_*\| \rightarrow d$, where $\|d\| = 1$. Denoting $t_k := \|x_k - x_*\|$, one has

$$x_k = x_* + t_k d + o(t_k).$$

Let us show that d is the desired critical direction.

- On the one hand, because Θ_σ is Lipschitz continuous near x_* , the following holds

$$\Theta_\sigma(x_k) = \Theta_\sigma(x_* + t_k d) + o(t_k).$$

This estimate and (5.20) show that $\Theta'_\sigma(x_*; d) \leq 0$. Then, from lemma 5.9, one can write

$$f'(x_*) \cdot d + \sigma \|P_{c(x_*)}(c'(x_*) \cdot d)\|_{\mathbb{P}} \leq 0. \quad (5.21)$$

This certainly implies that

$$f'(x_*) \cdot d \leq 0. \quad (5.22)$$

- On the other hand, from the assumptions, there is an optimal multiplier $\hat{\lambda}_*$ such that $\sigma > \|\hat{\lambda}_*\|_{\mathbb{D}}$. We have

$$\begin{aligned} -f'(x_*) \cdot d &= \hat{\lambda}_*^{\top} (c'(x_*) \cdot d) \quad [\nabla_x \ell(x_*, \hat{\lambda}_*) = 0] \\ &\leq \hat{\lambda}_*^{\top} P_{c(x_*)}(c'(x_*) \cdot d) \quad [(\hat{\lambda}_*)_I \geq 0 \text{ and } (\hat{\lambda}_*)_I^{\top} c_I(x_*) = 0] \\ &\leq \|\hat{\lambda}_*\|_{\mathbb{D}} \|P_{c(x_*)}(c'(x_*) \cdot d)\|_{\mathbb{P}} \quad [(5.18)]. \end{aligned}$$

Then (5.21) and $\sigma > \|\hat{\lambda}_*\|_{\mathbb{D}}$ imply that $P_{c(x_*)}(c'(x_*) \cdot d) = 0$, i.e.,

$$\begin{cases} c'_i(x_*) \cdot d = 0 & \text{for } i \in E \\ c'_i(x_*) \cdot d \leq 0 & \text{for } i \in I_*^0. \end{cases}$$

These and (5.22) show that d is a nonzero critical direction.

Now, let λ_* be the multiplier depending on d , determined by the weak second-order sufficient condition of optimality (2.80). By this condition and $d \in C_* \setminus \{0\}$, one has

$$\langle \nabla_{xx}^2 \ell(x_*, \lambda_*) d, d \rangle > 0.$$

The following Taylor expansion (use $\nabla_x \ell(x_*, \lambda_*) = 0$)

$$\ell(x_k, \lambda_*) = \ell(x_*, \lambda_*) + \frac{t_k^2}{2} \langle \nabla_{xx}^2 \ell(x_*, \lambda_*) d, d \rangle + o(t_k^2)$$

allows us to see that, for k large enough,

$$\ell(x_k, \lambda_*) > \ell(x_*, \lambda_*). \tag{5.23}$$

Then, for large indices k , there holds

$$\begin{aligned} \Theta_\sigma(x_k) &\leq \Theta_\sigma(x_*) && [(5.20)] \\ &= f(x_*) && [c(x_*)^\# = 0] \\ &= \ell(x_*, \lambda_*) && [\lambda_*^\top c(x_*) = 0] \\ &< \ell(x_k, \lambda_*) && [(5.23)] \\ &\leq \Theta_\sigma(x_k) && [\text{lemma 5.10, } \sigma \geq \|\lambda_*\|_D, \text{ and } (\lambda_*)_I \geq 0]. \end{aligned}$$

We have shown $\Theta_\sigma(x_k) < \Theta_\sigma(x_k)$, which is the expected contradiction. □

5.1.4 Globalization by Line-Search for (P_{EI}) ▲

Definition of the algorithm

The *global extension* of the *local SQP* algorithm 5.1 analyzed in this section replaces the Hessian of the Lagrangian $\nabla_{xx}^2 \ell(x_k, \lambda_k)$ by a positive definite linear operator $M_k : \mathbb{E} \rightarrow \mathbb{E}$ (a property that we condense by the notation $M_k > 0$). Hence, the *osculating quadratic problem* (5.9) becomes

$$(OQP)_k \quad \begin{cases} \min_d \langle \nabla f(x_k), d \rangle + \frac{1}{2} \langle M_k d, d \rangle \\ c_E(x_k) + c'_E(x_k) d = 0 \\ c_I(x_k) + c'_I(x_k) d \leq 0. \end{cases} \tag{5.24}$$

We shall need the KKT system of this optimization problem: if d_k solves (5.24), then there exists a multiplier $\lambda_k^{\text{QP}} \in \mathbb{R}^m$ (since the constraints are qualified by (CQ-A)) such that

$$\nabla f(x_k) + M_k d_k + c'(x_k)^* \lambda_k^{\text{QP}} = 0, \tag{5.25a}$$

$$c_E(x_k) + c'_E(x_k) d_k = 0, \tag{5.25b}$$

$$0 \leq (\lambda_k^{\text{QP}})_I \perp (c_I(x_k) + c'_I(x_k) d_k) \leq 0. \tag{5.25c}$$

There are several reasons motivating the choice of making the substitution $\nabla_{xx}^2 \ell(x_k, \lambda_k) \curvearrowright M_k$, with a positive definite operator M_k . Some are related to remark 5.2:

- problem $(OQP)_k$ has more often a solution; this is actually the case if and only if its constraints are compatible (Frank and Wolfe [56]); furthermore, there are techniques that can face the situations where the constraints are inconsistent (but this is more technical);
- when $(OQP)_k$ has a solution, this one is unique in d_k (but not in λ_k^{QP} , whose uniqueness depends on a constraint qualification);
- problem $(OQP)_k$ can be solved in polynomial time, which is crucial for the efficiency of the overall algorithm.

The inconvenient of this modification of the direction definition is that the primal-dual quadratic convergence of theorem 5.6 is normally lost (at least if the Hessian of the Lagrangian is not positive definite at the solution). Nevertheless, when M_k is updated by a quasi-Newton formula, the convergence is often superlinear and a very precise solution can be obtained in very few iterations (for quasi-Newton methods the number of iterations is often roughly proportional to the number of variables, while with second derivative computation this number is independent of the number of variables).

Another reason for making the substitution $\nabla_{xx}^2 \ell(x_k, \lambda_k) \curvearrowright M_k$, with a positive definite operator M_k , is that the SQP direction d_k , solution to the above quadratic optimization problem $(OQP)_k$, is then a descent direction of the exact penalty function Θ_σ at x_k , as claimed by the following proposition.

Proposition 5.12 (descent property of the SQP direction) *Suppose that f and c are differentiable at $x_k \in \mathbb{E}$, that $(d_k, \lambda_k^{QP}) \in \mathbb{E} \times \mathbb{R}^m$ is a stationary pair of the osculating quadratic optimization problem (5.24), and that $\|\cdot\|_P$ is convex. Then,*

1) $\Theta'_k(x_k; d_k) \leq \Delta_k$, where

$$\Delta_k := \langle \nabla f(x_k), d_k \rangle - \sigma \|c(x_k)^\# \|_P, \quad (5.26a)$$

$$= -\langle M_k d_k, d_k \rangle + (\lambda_k^{QP})^\top c(x_k) - \sigma \|c(x_k)^\# \|_P, \quad (5.26b)$$

$$\leq -\langle M_k d_k, d_k \rangle + (\|\lambda_k^{QP}\|_D - \sigma) \|c(x_k)^\# \|_P, \quad (5.26c)$$

2) if, furthermore, $\sigma \geq \|\lambda_k^{QP}\|_D$, then $\Theta'_k(x_k; d_k) \leq -\langle M_k d_k, d_k \rangle$,

3) if, furthermore, $M_k > 0$, then $\Theta'_k(x_k; d_k) \leq 0$,

4) if, furthermore, x_k is not a stationary point of (P_{EI}) , then $\Theta'_k(x_k; d_k) < 0$.

Proof. 1) Using lemma 5.8, we get

$$\Theta'(x_k; d_k) = \langle \nabla f(x_k), d_k \rangle + \sigma (\|\cdot\|_P)'(c(x_k); c'(x_k)d_k).$$

Let us examine the last directional derivative. One has

$$\begin{aligned}
 & (\|\cdot\|_{\mathbb{P}}^{\#})'(c(x_k); c'(x_k)d_k) \\
 &= \lim_{t \downarrow 0} \frac{1}{t} \left(\left\| \underbrace{c(x_k) + t c'(x_k)d_k}_{=(1-t)c(x_k) + t(c(x_k) + c'(x_k)d_k)} \right\|_{\mathbb{P}}^{\#} - \|c(x_k)\|_{\mathbb{P}}^{\#} \right) \\
 & \leq (1-t)\|c(x_k)\|_{\mathbb{P}}^{\#} + t\|(c(x_k) + c'(x_k)d_k)\|_{\mathbb{P}}^{\#} \\
 & \quad [\text{convexity of } \|\cdot\|_{\mathbb{P}}^{\#}] \\
 & \leq -\|c(x_k)\|_{\mathbb{P}}^{\#} \quad [(c(x_k) + c'(x_k)d_k)^{\#} = 0],
 \end{aligned}$$

where we have used the constraints of the OQP (5.24). This gives (5.26a). For getting (5.26b), we rewrite $\langle \nabla f(x_k), d_k \rangle$ as follows

$$\begin{aligned}
 \langle \nabla f(x_k), d_k \rangle &= -\langle M_k d_k, d_k \rangle - (\lambda_k^{\text{QP}})^{\text{T}} c'(x_k) d_k \quad [(5.25a)] \\
 &= -\langle M_k d_k, d_k \rangle + (\lambda_k^{\text{QP}})^{\text{T}} c(x_k),
 \end{aligned}$$

since $c'_E(x_k)d_k = -c_E(x_k)$ by (5.25b) and $(\lambda_k^{\text{QP}})^{\text{T}} c'_I(x_k)d_k = -(\lambda_k^{\text{QP}})^{\text{T}} c_I(x_k)$ by (5.25c). Inequality (5.26c) is now a consequence of (5.26b) and (5.18).

2) We have to show that, when $\sigma \geq \|\lambda_k^{\text{QP}}\|_{\text{D}}$, the last two terms in (5.26b) form a nonpositive difference. This is indeed the case, since

$$\begin{aligned}
 & (\lambda_k^{\text{QP}})^{\text{T}} c(x_k) - \sigma \|c(x_k)\|_{\mathbb{P}}^{\#} \\
 & \leq (\lambda_k^{\text{QP}})^{\text{T}} c(x_k)^{\#} - \sigma \|c(x_k)\|_{\mathbb{P}}^{\#} \quad [(\lambda_k^{\text{QP}})_I \geq 0] \\
 & \leq (\|\lambda_k^{\text{QP}}\|_{\text{D}} - \sigma) \|c(x_k)\|_{\mathbb{P}}^{\#} \quad [(5.18)] \\
 & \leq 0 \quad [\sigma \geq \|\lambda_k^{\text{QP}}\|_{\text{D}}].
 \end{aligned}$$

3) Clear.

4) We proceed by contraposition. If $\Theta'_k(x_k; d_k) = 0$, $d_k = 0$ by the inequality in point 2. Now (5.25) with $d_k = 0$ shows that $(x_k, \lambda_k^{\text{QP}})$ is a stationary pair of (PEI) . \square

The assumption on the convexity of $\|\cdot\|_{\mathbb{P}}^{\#}$ in the previous proposition is satisfied by many standard norm, in particular by the ℓ_p -norms with $1 \leq p \leq \infty$, but not by all the norms. This question is examined in exercise 5.1.2.

The sufficient condition guaranteeing the descent of Θ_{σ} along the SQP direction d_k , in point 2 of the previous proposition, namely $\sigma \geq \|\lambda_k^{\text{QP}}\|_{\text{D}}$, recalls the sufficient condition $\sigma > \sup\{\|\lambda_{*}\|_{\text{D}} : \lambda_{*} \in \Lambda\}$, exhibited by theorem 5.11, that ensures the exactness of the merit function Θ_{σ} . This is not surprising and reassuring on the correctness of the analysis.

Observe now that the threshold $\sup\{\|\lambda_{*}\|_{\text{D}} : \lambda_{*} \in \Lambda\}$ above which σ must be is not known by the algorithm or the user of the algorithm (it depends on the optimal multipliers that are not known) but, in a certain way, the values $\|\lambda_k^{\text{QP}}\|_{\text{D}}$, evolving along the iterations, can inform the algorithm on the value that σ must have to make Θ_{σ} exact at the limit point of the generated sequence of primal iterates $\{x_k\}$. This observation also indicates that the value of σ , which is monitored by the algorithm, may have to be updated during the iterative process. Therefore σ depend on the iteration counter k and is now denoted by σ_k . Specific update rules are used to ensure that

$$\sigma_k \geq \|\lambda_k^{\text{QP}}\|_{\text{D}} + \bar{\sigma}, \quad (5.27a)$$

where $\bar{\sigma} > 0$ is some constant safeguard. Typically, one takes

$$\bar{\sigma} := \max(\sqrt{\mathbf{e}}, \|\lambda_1^{\text{QP}}\|_{\text{D}}/100) \quad \text{and} \quad \sigma_1 := \|\lambda_1^{\text{QP}}\|_{\text{D}} + \bar{\sigma},$$

where \mathbf{e} is the *machine-epsilon*. The inequality (5.27a) does not prevent from taking arbitrarily σ_k very large, which is both harmful from the numerical efficiency of the algorithm and for its convergence analysis (for example, because one can force σ_k to blow up without a reason motivated by the behavior of the algorithm). For this reason, one also requires that the following condition be satisfied:

$$\{\lambda_k^{\text{QP}}\} \text{ is bounded} \implies \sigma_k \text{ is constant for } k \text{ large.} \quad (5.27b)$$

An example of update rule of σ_k satisfying these conditions (5.27) is the following.

Rule 5.13 (update of $\sigma_k - \text{I}$) Given a threshold $\bar{\sigma} > 0$, the current OQP multiplier λ_k^{QP} and the previous penalty parameter σ_{k-1} , compute the new penalty parameter σ_k as follows.

if $\sigma_{k-1} \geq \|\lambda_k^{\text{QP}}\|_{\text{D}} + \bar{\sigma}$;
then $\sigma_k = \sigma_{k-1}$;
else $\sigma_k = \max(1.5\sigma_{k-1}, \|\lambda_k^{\text{QP}}\|_{\text{D}} + \bar{\sigma})$;

This update rule has the nice property to eventually fix σ_k to a constant value when the sequence $\{\lambda_k^{\text{QP}}\}$ is bounded. It is not without flaw, however, since $\{\sigma_k\}$ is nondecreasing and is, therefore, penalized by a badly chosen initial penalty parameter σ_1 or a high value of the parameter determined far from the solution. For this reason, one often adds instructions to decrease σ_k if this one is clearly too large, with respect to the current $\|\lambda_k^{\text{QP}}\|_{\text{D}}$.

Rule 5.14 (update of $\sigma_k - \text{II}$) Given a threshold $\bar{\sigma} > 0$, the current OQP multiplier λ_k^{QP} and the previous penalty parameter σ_{k-1} , compute the new penalty parameter σ_k as follows.

if $\sigma_{k-1} \geq 1.1(\|\lambda_k^{\text{QP}}\|_{\text{D}} + \bar{\sigma})$;
then $\sigma_k = (\sigma_{k-1} + \|\lambda_k^{\text{QP}}\|_{\text{D}} + \bar{\sigma})/2$;
else
if $\sigma_{k-1} \geq \|\lambda_k^{\text{QP}}\|_{\text{D}} + \bar{\sigma}$;
then $\sigma_k = \sigma_{k-1}$;
else $\sigma_k = \max(1.5\sigma_{k-1}, \|\lambda_k^{\text{QP}}\|_{\text{D}} + \bar{\sigma})$;

With the rule 5.14, property (5.27b) is no longer guaranteed, however, and the convergence result of proposition 5.17 below is no longer ensured.

One can now state a frequently used line-search version of the SQP algorithm.

Algorithm 5.15 (SQP with line-search) Let $\beta \in (0, 1)$ and $\omega \in (0, \frac{1}{2})$ be parameters used in the line-search (typically $\beta = 1/2$ and $\omega = 10^{-4}$). One iteration, from $(x_k, \lambda_k, M_k) \in \mathbb{E} \times \mathbb{R}^m \times \mathbb{R}^{n \times n}$ to $(x_{k+1}, \lambda_{k+1}, M_{k+1})$ is made of the following steps (the penalty parameter σ_k is also updated):

1. *Stopping test*: if the current pair (x_k, λ_k) is satisfactory, stop;
2. *QP solve*: let $(d_k, \lambda_k^{\text{QP}})$ be the primal-dual solution to the osculating quadratic problem (5.24), if any;
3. *New penalty parameter*: update σ_k so that (5.27) is satisfied;
4. *Armijo line-search*: determine $\alpha_k := \beta^{i_k}$ where i_k is the smallest nonnegative integer such that

$$\Theta_{\sigma_k}(x_k + \alpha_k d_k) \leq \Theta_{\sigma_k}(x_k) + \omega \alpha_k \Delta_k, \quad (5.28)$$

where Δ_k is given by (5.26);

5. *Update M_k* : by some technique (for example, modification of $\nabla_{xx}^2 \ell(x_k, \lambda_k)$ or a quasi-Newton update);
6. *New iterate*: set $x_{k+1} := x_k + \alpha_k d_k$ and $\lambda_{k+1} = \lambda_k + \alpha_k (\lambda_k^{\text{QP}} - \lambda_k)$;

Remarks 5.16 1) The algorithm cannot do better than finding a stationary point of problem (PEI), in particular because when it starts at such a point it may compute a vanishing displacement d_k (and will do so if $M_1 > 0$). Therefore, the only reasonable stopping criterion in step 1 is to test the approximate satisfaction of the first order (KKT) optimality conditions (5.1).

2) In the line-search of step 4, instead of giving to the trial stepsizes the predetermined values β^{i_k} it is better to do interpolation.

Global convergence

Proposition 5.17 (global convergence of the SQP algorithm with line-search) Suppose that f and c are $\mathcal{C}^{1,1}$, that $\|\cdot\|_{\text{P}}^{\#}$ is convex, that the sequences $\{M_k\}$ and $\{M_k^{-1}\}$ are bounded, that at each iteration the OQP (5.24) has a solution $(d_k, \lambda_k^{\text{QP}})$, that the update rule of σ_k satisfies (5.27), and that $\Theta_k(x_k)$ is bounded below. Then, one of the following two complementary situations occurs

- 1) the sequence $\{\sigma_k\}$ is unbounded, in which case $\{\lambda_k^{\text{QP}}\}$ is also unbounded,
- 2) the sequence $\{\sigma_k\}$ is bounded, in which case algorithm 5.15 converges, in the sense that

$$\nabla_x \ell(x_k, \lambda_k^{\text{QP}}) \rightarrow 0, \quad c(x_k)^{\#} \rightarrow 0, \quad (\lambda_k^{\text{QP}})_I \geq 0 \quad \text{and} \quad (\lambda_k^{\text{QP}})_I^{\text{T}} c_I(x_k) \rightarrow 0.$$

Proof. 1) If $\{\sigma_k\}$ is unbounded, we see from rule (5.27b) that $\{\lambda_k^{\text{QP}}\}$ is unbounded.

2) If $\{\sigma_k\}$ is bounded, the inequality (5.27a) shows that $\{\lambda_k^{\text{QP}}\}$ is also bounded. Then, the rule (5.27b) guarantees that there exists an index k_1 such that $\sigma_k = \sigma$ for

all $k \geq k_1$. Now, each iteration after k_1 forces the decrease in the same function Θ_σ . Since $\{\Theta_\sigma(x_k)\}$ is bounded below, Armijo's condition (5.28) shows that

$$\alpha_k \Delta_k \rightarrow 0.$$

If we show $\{\alpha_k\}$ is bounded away from zero (i.e., $\alpha_k \geq \underline{\alpha}$ for some $\underline{\alpha} > 0$), the result follows. Indeed, then $\Delta_k \rightarrow 0$, and with (5.26c) and (5.27a), we deduce that

$$d_k^\top M_k d_k \rightarrow 0 \quad \text{and} \quad c(x_k)^\# \rightarrow 0,$$

so that the second claim is satisfied. Because M_k is positive definite and has a bounded inverse, $d_k \rightarrow 0$. Then, from (5.25a) and the boundedness of M_k , we see that $\nabla_x \ell(x_k, \lambda_k^{\text{QP}}) \rightarrow 0$, so that the first claim is satisfied. Furthermore, (5.25c) shows that $(\lambda_k^{\text{QP}})_I \geq 0$, so that the third claim is satisfied. Finally, $\Delta_k = \nabla f(x_k)^\top d_k - \sigma \|c(x_k)^\#\|_{\mathbb{P}} \rightarrow 0$ and $c(x_k)^\# \rightarrow 0$ imply that $\nabla f(x_k)^\top d_k \rightarrow 0$ and, using (5.25a), $(\lambda_k^{\text{QP}})^\top c'(x_k) d_k \rightarrow 0$. Hence

$$\begin{aligned} (\lambda_k^{\text{QP}})_I^\top c_I(x_k) &= -(\lambda_k^{\text{QP}})_I^\top c'_I(x_k) d_k && [(5.25c)] \\ &= (\lambda_k^{\text{QP}})_E^\top c'_E(x_k) d_k + o(1) && [(\lambda_k^{\text{QP}})^\top c'(x_k) d_k \rightarrow 0] \\ &= -(\lambda_k^{\text{QP}})_E^\top c_E(x_k) + o(1) && [(5.25b)] \\ &= o(1), \end{aligned}$$

because $\{\lambda_k^{\text{QP}}\}$ is bounded and $c_E(x_k) \rightarrow 0$. We have shown the fourth and last claim.

Therefore, it remains to prove that $\alpha_k \geq \underline{\alpha} > 0$, for all k and some constant $\underline{\alpha}$, which is a rather technical part of the proof. We can consider the indices k of $\mathcal{K} := \{k \geq k_1 : \alpha_k < 1\}$. Then, from the rule determining the stepsize α_k , the Armijo inequality is not verified for $\bar{\alpha}_k := \alpha_k/\beta$:

$$\Theta_\sigma(x_k + \bar{\alpha}_k d_k) > \Theta_\sigma(x_k) + \omega \bar{\alpha}_k \Delta_k. \quad (5.29)$$

Let us expand the left-hand side of (5.29). Using the smoothness of f and c , $\bar{\alpha}_k \leq 1$, the convexity of $\|\cdot\|_{\mathbb{P}}^\#$ (hence its Lipschitz continuity), (5.25b), and finally (5.26), we have successively

$$\begin{aligned} f(x_k + \bar{\alpha}_k d_k) &= f(x_k) + \bar{\alpha}_k \nabla f(x_k)^\top d_k + O(\bar{\alpha}_k^2 \|d_k\|^2) \\ c(x_k + \bar{\alpha}_k d_k) &= c(x_k) + \bar{\alpha}_k c'(x_k) d_k + O(\bar{\alpha}_k^2 \|d_k\|^2) \\ &= (1 - \bar{\alpha}_k) c(x_k) + \bar{\alpha}_k (c(x_k) + c'(x_k) d_k) + O(\bar{\alpha}_k^2 \|d_k\|^2) \\ \|c(x_k + \bar{\alpha}_k d_k)^\#\|_{\mathbb{P}} &\leq (1 - \bar{\alpha}_k) \|c(x_k)^\#\|_{\mathbb{P}} + \bar{\alpha}_k \|c(x_k) + c'(x_k) d_k\|_{\mathbb{P}}^\# + O(\bar{\alpha}_k^2 \|d_k\|^2) \\ &= (1 - \bar{\alpha}_k) \|c(x_k)^\#\|_{\mathbb{P}} + O(\bar{\alpha}_k^2 \|d_k\|^2) \\ \Theta_\sigma(x_k + \bar{\alpha}_k d_k) &\leq \Theta_\sigma(x_k) + \bar{\alpha}_k \Delta_k + C_1 \bar{\alpha}_k^2 \|d_k\|^2. \end{aligned}$$

Then (5.29) yields

$$-(1 - \omega) \bar{\alpha}_k \Delta_k \leq C_1 \bar{\alpha}_k^2 \|d_k\|^2.$$

But $\Delta_k = -d_k^\top M_k d_k + (\lambda_k^{\text{QP}})^\top c(x_k) - \sigma \|c(x_k)^\#\|_{\mathbb{P}} \leq -d_k^\top M_k d_k \leq -C_2 \|d_k\|^2$ (boundedness of $\{M_k^{-1}\}$), so that we deduce from the above inequality:

$$\bar{\alpha}_k \geq (C_2/C_1)(1 - \omega) > 0,$$

because $\omega < 1$. The positive lower bound on α_k can therefore be taken as $\underline{\alpha} := \beta(C_2/C_1)(1 - \omega)$. This concludes the proof. \square

Admissibility of the unit stepsize ▲

5.1.5 Globalization by Trust-Region for (P_E) ▲

Notes ▲

The analysis of the local convergence of the SQP algorithm, inherited from the one of the JN algorithm (section 5.1.2), has been taken up from [19; 1994]. This one gives the local convergence result of theorem 5.6, with the weakest assumptions on the limit point (uniqueness of the optimal multiplier and SC2) known so far. Before that, one had results assuming the SC2 and the stronger (CQ-LI) [20; theorem 15.4] and sometimes strict complementarity ([17; p. 252-256], [20; theorem 15.2], and the original work [117]).

For the algorithmic issues, we refer the reader to the state of the art by Fletcher [55].

Exercises

5.1.1. Nondifferentiable augmented Lagrangian. Let \mathbb{E} be a Euclidean vector space. Consider the optimization problem (P_{EI}) and its Lagrangian $\ell : (x, \lambda) \in \mathbb{E} \times \mathbb{R}^m \rightarrow \ell(x, \lambda) = f(x) + \lambda^\top c(x) \in \mathbb{R}$. Let x_* be a local minimum of problem (P_{EI}) at which the KKT conditions hold and denote by Λ_* the set of optimal multipliers associated with x_* . Suppose that the weak second-order sufficient condition of optimality holds at x_* . Let $\|\cdot\|_p$ be a norm on \mathbb{R}^m and $\|\cdot\|_D$ be its dual norm with respect to the Euclidean scalar product. Let $\mu \in \mathbb{R}^m$ and $\sigma \in \mathbb{R}_+$ verifying

$$\sigma \geq \sup_{\lambda_* \in \Lambda_*} \|\lambda_* - \mu\|_D \quad \text{and} \quad \sigma > \|\hat{\lambda}_* - \mu\|_D, \text{ for some } \hat{\lambda}_* \in \Lambda_*. \quad (5.30)$$

We want to show that the function $\Theta_{\mu, \sigma} : \mathbb{E} \rightarrow \mathbb{R}$ defined at $x \in \mathbb{E}$ by

$$\Theta_{\mu, \sigma}(x) := f(x) + \mu^\top c(x)^\# + \sigma \|c(x)^\#\|_p$$

has a strict local minimum at x_* . We propose a reasoning by contradiction.

1. Show that if $\Theta_{\mu, \sigma}$ has not a strict local minimum at x_* , one can find a sequence $\{x_k\} \subseteq \mathbb{E}$, a sequence of positive real numbers $\{t_k\} \downarrow 0$, and a nonzero critical direction d such that $x_k = x_* + t_k d + o(t_k)$.
2. Show that

$$\exists \lambda_* \in \Lambda_*, \quad \forall k \text{ large} : \quad \ell(x_*, \lambda_*) < \ell(x_k, \lambda_*).$$
3. Get a contradiction.

5.1.2. Norm assumptions. For an arbitrary norm $\|\cdot\|$ on \mathbb{R}^m , show that the following properties are equivalent (the operators $|\cdot|$ and $(\cdot)^+$ act componentwise):

- (i) $0 \leq u \leq v \Rightarrow \|u\| \leq \|v\|$,
- (ii) $u \leq v \Rightarrow \|u^+\| \leq \|v^+\|$,
- (iii) $v \mapsto \|v^+\|$ is convex.

Remark: These equivalences show that the convexity of $\|\cdot\|_p$ assumed in proposition 5.12 is satisfied with the ℓ_p norms, $1 \leq p \leq \infty$, since the ℓ_p norms satisfy (i).

5.2 SQP Algorithm for (P_G) ▲

5.2.1 The SQP Algorithm

Consider the optimization problem (P_G) in (2.1) with its Lagrangian $\ell : \mathbb{E} \times \mathbb{F} \rightarrow \mathbb{R}$ defined at $(x, \lambda) \in \mathbb{E} \times \mathbb{F}$ by

$$\ell(x, \lambda) = f(x) + \langle \lambda, c(x) \rangle. \quad (5.31)$$

The first order optimality conditions (2.8) at a primal-dual solution (x, λ) (we drop the star indices to alleviate notation) read

$$\nabla f(x) + c'(x)^* \lambda = 0 \quad \text{and} \quad \lambda \in N_G(c(x)). \quad (5.32)$$

Let us show that this system reads like the following functional inclusion

$$F(z) + N_C(z) \ni 0, \quad (5.33)$$

for some variable z , some function F , and some closed convex set C . Observe first that (5.32) also reads as the following system in $z := (x, y, \lambda) \in \mathbb{E} \times \mathbb{F} \times \mathbb{F}$:

$$\nabla f(x) + c'(x)^* \lambda = 0, \quad \lambda \in N_G(y) \quad \text{and} \quad y - c(x) = 0. \quad (5.34)$$

This one can be written like (5.33) in the variable z , where the function F and the closed convex set C are defined by

$$F(z) = \begin{pmatrix} \nabla f(x) + c'(x)^* \lambda \\ -\lambda \\ y - c(x) \end{pmatrix} \quad \text{and} \quad C = \mathbb{E} \times G \times \mathbb{F}. \quad (5.35)$$

Indeed, $N_C(z) = N_{\mathbb{E}}(x) \times N_G(y) \times N_{\mathbb{F}}(y) = \{0_{\mathbb{E}}\} \times N_G(y) \times \{0_{\mathbb{F}}\}$, so that (5.34) and (5.33) are the same systems.

The Josephy-Newton (JN) algorithm consists in determining $z_+ := (x_+, y_+, \lambda_+)$ as a solution to the linearized functional inclusion

$$F(z) + F'(z) \cdot (z_+ - z) + N_C(z_+) \ni 0.$$

We have

$$F'(z) = \begin{pmatrix} L & 0 & c'(x)^* \\ 0 & 0 & -I \\ -c'(x) & I & 0 \end{pmatrix} \quad (5.36)$$

where $L := \nabla_{xx}^2 \ell(x, \lambda)$. Therefore, z_+ is determined as a solution to

$$\begin{cases} \nabla f(x) + c'(x)^* \lambda + L(x_+ - x) + c'(x)^* (\lambda_+ - \lambda) = 0 \\ \lambda + (\lambda_+ - \lambda) \in N_G(y_+) \\ y - c(x) - c'(x)(x_+ - x) + (y_+ - y) = 0, \end{cases}$$

which after elimination of y and λ becomes

$$\begin{cases} \nabla f(x) + L(x_+ - x) + c'(x)^* \lambda_+ = 0 \\ \lambda_+ \in N_G(y_+) \\ c(x) + c'(x)(x_+ - x) = y_+. \end{cases}$$

After elimination of y_+ and the introduction of $d = x_+ - x$, we get

$$\begin{cases} \nabla f(x) + Ld + c'(x)^* \lambda_+ = 0 \\ \lambda_+ \in N_G(c(x) + c'(x)d). \end{cases}$$

This system is the first order optimality conditions of the *osculating quadratic problem*

$$\text{(OQP)} \quad \begin{cases} \min_d \langle \nabla f(x), d \rangle + \frac{1}{2} \langle \nabla_{xx}^2 \ell(x, \lambda) d, d \rangle \\ c(x) + c'(x) \cdot d \in G. \end{cases} \quad (5.37)$$

Algorithm 5.18 (local SQP for (P_G)) One iteration, from $(x_k, \lambda_k) \in \mathbb{E} \times \mathbb{F}$ to $(x_{k+1}, \lambda_{k+1}) \in \mathbb{E} \times \mathbb{F}$ is made of the following steps:

1. *Stopping test*: if the current pair (x_k, λ_k) is satisfactory, stop;
2. *QP solve*: let $(d_k, \lambda_k^{\text{QP}})$ be an *appropriate* primal-dual solution to the osculating quadratic problem (5.37), if any;
3. *New iterate*: set $x_{k+1} := x_k + d_k$ and $\lambda_{k+1} := \lambda_k^{\text{QP}}$.

5.2.2 Local Convergence

6 Self-dual Conic Optimization

Let \mathbb{E} and \mathbb{F} be two finite dimension vector spaces, and $\langle \cdot, \cdot \rangle$ be a scalar product on \mathbb{E} . A *conic optimization problem* is an optimization problem of the form

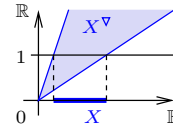
$$\begin{cases} \inf_{x \in \mathbb{E}} \langle c, x \rangle \\ \mathcal{A}(x) = b \\ x \in K, \end{cases} \quad (6.1)$$

where $c \in \mathbb{E}$, $\mathcal{A} : \mathbb{E} \rightarrow \mathbb{F}$ is a linear map, $b \in \mathbb{F}$, and K is a closed cone of \mathbb{E} . Hence, one minimizes a linear function on the intersection of an affine space and a closed cone.

Problem (6.1) may look like a very structured problem, but without more specifications on K , a problem as general as $\inf_x \{f(x) : x \in X\}$, where $f : \mathbb{E} \rightarrow \mathbb{R}$ and $X \subseteq \mathbb{E}$, can be written in the form (6.1). This is a consequence of the following two observations.

- There is no restriction in assuming that the objective of an optimization problem is linear, like in (6.1). For example, the problem $\inf_{x \in \mathbb{E}} \{f(x) : x \in X\}$ can be rewritten into the problem $\inf_{(x,t) \in \mathbb{E} \times \mathbb{R}} \{t : f(x) \leq t, x \in X\}$ (the equivalence is about the optimal values and the x -part of the solutions).
- We are now reduced to rewrite a problem of the form $\inf_{x \in \mathbb{E}} \{\langle c, x \rangle : x \in X\}$ into the form (6.1). This is done by a technique called *conification* of X (also called *homogenization* [125; § 8], [118], but we prefer to avoid this term, which is used in the field of partial differential equations with a completely different meaning). One introduces the closed cone X^∇ of $\mathbb{E} \times \mathbb{R}$ defined by

$$\begin{aligned} X^\nabla &:= \text{cl } X^\vee, \quad \text{where} \\ X^\vee &:= \{(x, \alpha) \in \mathbb{E} \times \mathbb{R} : \alpha > 0, \alpha^{-1}x \in X\}. \end{aligned}$$



Since X^\vee is clearly a cone of $\mathbb{E} \times \mathbb{R}$, X^∇ is a closed cone. Furthermore, $x \in X$ if and only if $(x, 1) \in X^\nabla$. Therefore, the problem $\inf_{x \in \mathbb{E}} \{\langle c, x \rangle : x \in X\}$ also reads $\inf_{(x,\alpha) \in \mathbb{E} \times \mathbb{R}} \{\langle c, x \rangle : (x, \alpha) \in X^\nabla, \text{ and } \alpha = 1\}$, which is in the form of (6.1).

Therefore, to really take advantage of the structure of (6.1), more assumptions must be given on the cone K .

A certainly much more specific problem is obtained with (6.1) if one assumes that K is a closed *convex* cone. This does not imply that problem (6.1) becomes easy to understand and to solve. A rather fruitful analysis is possible at this level of generality [14, 116], but in this chapter will limit our presentation to three particular

self-dual cones K , meaning that they verify $K^+ = K$ (hence K is necessarily a closed convex cone).

- A *semidefinite optimization* problem is obtained from (6.1) when $\mathbb{E} = \mathcal{S}^n$ is the space of symmetric matrices of order n and $K = \mathcal{S}_+^n$ is the cone of positive semidefinite matrices. This problem is analyzed in section 6.1.
- A *circular optimization* problem arises when $\mathbb{E} = \mathbb{R}^n$ and $K = \mathbb{R}_\nabla^n$ is the circular cone (section 6.2).
- In *copositive optimization* (section 6.3), the cone is formed of symmetric matrices ($\mathbb{E} = \mathcal{S}^n$ again) whose associated quadratic form is nonnegative on the positive orthant.

These problems have been meticulously analyzed in the last decades for various reasons. The first reason deals with complexity issues. The first two problems can indeed be solved in a time that is polynomial in the dimension of the problem. The last one is NP-hard, despite the convexity of its objective, which is linear, and its convex feasible set; its computational complexity then comes from the complexity of the boundary of its feasible set, but the convexity of the latter and the convexity of the solution set make these problems very attractive to study. The second reason comes from the fact that many problems can be formulated as one of these three conic problems. If they can be written like the first two, they can then be solved in polynomial time. If the problem can be formulated as a copositive optimization problem, they benefit from the convexity of this problem, which makes it possible to approximate them by sequences of easier problems. As a last reason, let us mention that many nonconvex NP-hard problems have semidefinite or circular relaxations. Sometimes, this provides a rather precise lower bound on their optimal value, which is an important information in some branching approach to solve them.

6.1 Semidefinite Optimization

6.1.1 Primal and Dual Problems

The Cones \mathcal{S}_+^n and \mathcal{S}_{++}^n

Semidefinite optimization has an unknown belonging to the cone of positive definite matrices. We denote by \mathcal{S}^n the Euclidean space of symmetric $n \times n$ real matrices, equipped with the [scalar product](#)

$$\langle \cdot, \cdot \rangle : (A, B) \in (\mathcal{S}^n)^2 \mapsto \langle A, B \rangle = \text{tr}(AB) = \sum_{ij} A_{ij} B_{ij} \in \mathbb{R},$$

where $\text{tr} M := \sum_{i=1}^n M_{ii}$ denotes the *trace* of a square matrix M . It is easy to see that, for A and $B \in \mathbb{R}^{n \times n}$, $\text{tr} AB = \text{tr} BA$.

A real symmetric matrix $A \in \mathcal{S}^n$ has n real eigenvalues $\lambda_i \in \mathbb{R}$ with which are associated orthonormal eigenvectors v_i , $i \in [1 : n]$:

$$Av_i = \lambda_i v_i \quad \text{and} \quad v_i^\top v_j = \delta_{ij},$$

where δ_{ij} is the *Kronecker symbol* ($\delta_{ij} = 1$ if $i = j$, $\delta_{ij} = 0$ if $i \neq j$). If we denote by $V \in \mathbb{R}^{n \times n}$ the matrix whose columns are the eigenvectors v_1, \dots, v_n , we have

$AV = VA$, where $A = \text{Diag}(\lambda_1, \dots, \lambda_n)$. Since V is orthogonal, $V^{-1} = V^T$ and we get the *spectral decomposition* of $A \in \mathcal{S}^n$:

$$A = VAV^T = \sum_{i=1}^n \lambda_i v_i v_i^T.$$

The cone of \mathcal{S}^n made of the **positive semidefinite matrices** is denoted by \mathcal{S}_+^n . We use the notation $A \succeq 0$ to quote the fact that $A \in \mathcal{S}_+^n$. Recall that $A \succeq 0$ if its eigenvalues are nonnegative:

$$A \succeq 0 \iff A \in \mathcal{S}_+^n \iff \forall v \in \mathbb{R}^n : v^T A v \geq 0 \iff \lambda(A) \geq 0,$$

where we have denoted by $\lambda(A)$ the vector of the eigenvalues of $A \in \mathcal{S}^n$. The cone of \mathcal{S}^n made of the **positive definite matrices** is denoted by \mathcal{S}_{++}^n . We use the notation $A \succ 0$ to quote the fact that $A \in \mathcal{S}_{++}^n$. Recall that $A \succ 0$ if its eigenvalues are positive:

$$A \succ 0 \iff A \in \mathcal{S}_{++}^n \iff \forall v \in \mathbb{R}^n \setminus \{0\} : v^T A v > 0 \iff \lambda(A) > 0.$$

Lemma 6.1 (cones \mathcal{S}_+^n and \mathcal{S}_{++}^n)

- 1) $A \succeq 0 \iff \forall B \in \mathcal{S}_+^n : \langle A, B \rangle \geq 0$.
- 2) $A \succ 0 \iff \forall B \in \mathcal{S}_+^n \setminus \{0\} : \langle A, B \rangle > 0$.
- 3) For A and $B \in \mathcal{S}_+^n : \langle A, B \rangle = 0 \iff AB = 0$.

Proof. 1) Let A and $B \in \mathcal{S}_+^n$. Take $B = \sum_i \lambda_i v_i v_i^T$ the spectral decomposition of B . Then,

$$\langle A, B \rangle = \sum_{i=1}^n \lambda_i \langle A, v_i v_i^T \rangle = \sum_{i=1}^n \lambda_i \text{tr}(A v_i v_i^T) = \sum_{i=1}^n \lambda_i (v_i^T A v_i) \geq 0,$$

where we have used $\text{tr}(A v_i v_i^T) = \text{tr}(v_i^T A v_i) = v_i^T A v_i \in \mathbb{R}$ and, finally, the fact that $v_i^T A v_i \geq 0$ and $\lambda_i \geq 0$ by the positive semidefiniteness of A and B respectively.

Conversely, by taking $B = v v^T \succeq 0$ for an arbitrary $v \in \mathbb{R}^n$, we get that $0 \leq \langle A, v v^T \rangle = v^T A v$, which characterizes the positive semidefiniteness of A .

2) Let $A \succ 0$ and $B \succeq 0$ with $B \neq 0$. In the **spectral decomposition** of $B = \sum_{i=1}^n \lambda_i v_i v_i^T$, at least one of the λ_i 's is positive and the others are nonnegative. Then, like above, $\langle A, B \rangle = \sum_{i=1}^n \lambda_i (v_i^T A v_i)$, which is positive since all the $v_i^T A v_i > 0$ and all the $\lambda_i > 0$, with at least one $\lambda_i > 0$.

Conversely, assume that $\langle A, B \rangle > 0$ for all nonzero $B \succeq 0$. By taking $B = v v^T \succeq 0$, with an arbitrary nonzero v , one must have $0 < \langle A, B \rangle = v^T A v$, which shows that $A \succ 0$.

3) With the **spectral decomposition** of $B = \sum_{i=1}^n \lambda_i v_i v_i^T$ (the λ_i 's are ≥ 0) and $\langle A, B \rangle = 0$, one gets $\sum_i \lambda_i v_i A v_i^T = 0$. Since $A \in \mathcal{S}_+^n$, this implies that $A v_i = 0$ when $\lambda_i > 0$. Then, $AB = \sum_{i=1}^n \lambda_i A v_i v_i^T = 0$.

Conversely, it is clear that $AB = 0$ implies $\langle A, B \rangle = 0$. □

Here are some remarks and properties of the cones \mathcal{S}_+^n and \mathcal{S}_{++}^n .

- Point 1 of lemma 6.1 expresses that fact that \mathcal{S}_+^n is self-dual:

$$(\mathcal{S}_+^n)^+ = \mathcal{S}_+^n.$$

- Point 2 of lemma 6.1 is often used in the form of the following implication

$$\langle A, B \rangle = 0, \quad A > 0 \quad \text{and} \quad B \succeq 0 \quad \implies \quad B = 0. \quad (6.2)$$

- The *tangent* and *normal cone* to \mathcal{S}_+^n have the following expressions:

$$\mathcal{T}_A \mathcal{S}_+^n = \{D \in \mathcal{S}^n : v^T D v \geq 0, \text{ for all } v \in \mathcal{N}(A)\}, \quad (6.3)$$

$$\mathcal{N}_A \mathcal{S}_+^n = \{N \in \mathcal{S}_-^n : \langle A, N \rangle = 0\} = \mathcal{S}_-^n \cap \{A\}^\perp. \quad (6.4)$$

- \mathcal{S}_{++}^n is the cone of \mathcal{S}^n made of the *positive definite matrices*:

$$A \succeq 0 \text{ and } [v^T M v > 0, \forall v \in \mathcal{N}(A) \setminus \{0\}] \implies M + rA > 0 \text{ for large } r.$$

This last property is the so-called *Finsler lemma*.

Problem Definitions

The primal and (Lagrange) dual of the SDO problem read

$$(P) \quad \begin{cases} \inf_{X \in \mathcal{S}^n} \langle C, X \rangle \\ \mathcal{A}(X) = b \\ X \succeq 0 \end{cases} \quad \text{and} \quad (D) \quad \begin{cases} \sup_{(y, S) \in \mathbb{R}^m \times \mathcal{S}^n} b^T y \\ \mathcal{A}^*(y) + S = C \\ S \succeq 0, \end{cases} \quad (6.5)$$

where

- $C \in \mathcal{S}^n$ and $b \in \mathbb{R}^m$,
- $\mathcal{A} : \mathcal{S}^n \rightarrow \mathbb{R}^m$ is linear ($\mathcal{A}^* : \mathbb{R}^m \rightarrow \mathcal{S}^n$ is its adjoint).

Some notation

- Feasible sets:

$$\begin{aligned} \mathcal{F}_P &:= \{X \in \mathcal{S}_+^n : \mathcal{A}(X) = b\}, \\ \mathcal{F}_D &:= \{(y, S) \in \mathbb{R}^m \times \mathcal{S}_+^n : \mathcal{A}^*(y) + S = C\}, \\ \mathcal{F} &:= \mathcal{F}_P \times \mathcal{F}_D. \end{aligned}$$

- Strictly feasible sets:

$$\begin{aligned} \mathcal{F}_P^s &:= \{X \in \mathcal{S}_{++}^n : \mathcal{A}(X) = b\}, \\ \mathcal{F}_D^s &:= \{(y, S) \in \mathbb{R}^m \times \mathcal{S}_{++}^n : \mathcal{A}^*(y) + S = C\}, \\ \mathcal{F}^s &:= \mathcal{F}_P^s \times \mathcal{F}_D^s. \end{aligned}$$

- Optimal values: $\text{val}(P)$ and $\text{val}(D)$. Duality gap: $\text{val}(P) - \text{val}(D)$
- Solution sets: $\text{Sol}(P)$ and $\text{Sol}(D)$.

One can represent \mathcal{A} by m matrices $A_k \in \mathcal{S}^n$ (Riesz-Fréchet representation theorem), as follows

$$\mathcal{A}(X) = \begin{pmatrix} \langle A_1, X \rangle \\ \vdots \\ \langle A_m, X \rangle \end{pmatrix}. \quad (6.6a)$$

In this representation, the map \mathcal{A} is surjective if and only if the matrices A_k are linearly independent in \mathcal{S}^n . With the representation (6.6a) and the Euclidean scalar product on \mathbb{R}^m , the adjoint \mathcal{A}^* of \mathcal{A} takes at $y \in \mathbb{R}^m$ the value

$$\mathcal{A}^*(y) = \sum_{k \in [1:m]} y_k A_k \in \mathcal{S}^n. \quad (6.6b)$$

Concrete problems usually specify \mathcal{A} (resp. \mathcal{A}^*) using the representation (6.6a) (resp. (6.6b)), but the theory given in this chapter is easier to develop and present with the abstract forms (P) and (D) of the primal and dual problems.

The *Lagrangian* of problem (P) is the function $\ell : \mathcal{S}^n \times \mathbb{R}^m \times \mathcal{S}^n \rightarrow \mathbb{R}$ that takes at $(X, y, S) \in \mathcal{S}^n \times \mathbb{R}^m \times \mathcal{S}^n$ the value

$$\ell(X, y, S) = \langle C, X \rangle - \langle y, \mathcal{A}(X) - b \rangle - \langle S, X \rangle.$$

Note that the dual problem can also be written by eliminating the variable S as follows

$$(D) \quad \begin{cases} \sup_{y \in \mathbb{R}^m} b^\top y \\ C - \mathcal{A}^*(y) \succeq 0. \end{cases}$$

In that case it is more appropriate to take as Lagrangian of (P) the function that only dualises the equality constraint:

$$(X, y) \in \mathcal{S}^n \times \mathbb{R}^m \mapsto \langle C, X \rangle - \langle y, \mathcal{A}(X) - b \rangle.$$

Proposition 6.2 (consequences of the Lagrangian dualization)

- 1) $\text{val}(D) \leq \text{val}(P)$.
- 2) $(X, y, S) \in \mathcal{F} \Rightarrow \langle C, X \rangle - b^\top y = \langle X, S \rangle \geq 0$.
- 3) $(X, y, S) \in \mathcal{F}, \langle X, S \rangle = 0$
 $\Leftrightarrow X \in \text{Sol}(P), (y, S) \in \text{Sol}(D), \text{val}(D) = \text{val}(P),$
 $\Leftrightarrow (X, (y, S))$ is a saddle-point of ℓ on $\mathcal{S}^n \times (\mathbb{R} \times \mathcal{S}_+^n)$.

Proof. 1) This is the weak duality inequality (1.60).

2) If $(X, y, S) \in \mathcal{F}$, then $\langle C, X \rangle = \langle \mathcal{A}^*(y) + S, X \rangle = \langle y, \mathcal{A}(X) \rangle + \langle X, S \rangle = \langle y, b \rangle + \langle X, S \rangle$. Furthermore $\langle X, S \rangle \geq 0$ since $X \succeq 0$ and $S \succeq 0$.

3) [\Rightarrow] If $(X, y, S) \in \mathcal{F}$, the following holds

$$b^\top y \leq \text{val}(D) \leq \text{val}(P) \leq \langle C, X \rangle.$$

Since $\langle X, S \rangle = 0$, $b^\top y = \langle C, X \rangle$ by point 2 and therefore equality holds everywhere above, which implies that $X \in \text{Sol}(P)$, $(y, S) \in \text{Sol}(D)$, and $\text{val}(D) = \text{val}(P)$.

[\Leftarrow] Clearly, $(X, y, S) \in \mathcal{F}$, when $X \in \text{Sol}(P)$ and $(y, S) \in \text{Sol}(D)$. Furthermore, by point 2, $\langle X, S \rangle = \langle C, X \rangle - b^\top y = \text{val}(P) - \text{val}(D) = 0$. \square

6.1.2 Examples of SDO Formulation

Suppose that the principal submatrix A of the square symmetric matrix

$$K := \begin{pmatrix} A & B \\ B^\top & C \end{pmatrix}$$

is nonsingular. Then, the **Schur complement** of A in K is the matrix denoted and defined by

$$(A|K) := C - B^\top A^{-1} B.$$

This matrix can also be defined when K is nonsymmetric. It occurs in many circumstances [38]. From the block Gaussian factorization of K , which reads

$$K = \begin{pmatrix} I & 0 \\ B^\top A^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & (A|K) \end{pmatrix} \begin{pmatrix} I & A^{-1} B \\ 0 & I \end{pmatrix},$$

we deduce that the following characterization of the positive definiteness of K :

$$K > 0 \iff \begin{cases} A > 0 \\ (A|K) > 0. \end{cases} \quad (6.7)$$

As we shall see, this characterization often occurs in semi-definite optimization. For a characterization of the positive semi-definiteness of K with a Schur-like complement, see exercise 6.1.1.

Linear Optimization

Convex Quadratic Optimization

Global Minimization of Polynomials

Consider the problem of finding the *global minimum* of a real polynomial $p \in \mathbb{R}[x]$, in the variable $x \in \mathbb{R}^n$, which reads

$$\inf_{x \in \mathbb{R}^n} \left(p(x) := \sum_{\alpha \in \mathbb{N}^n} p_\alpha x^\alpha \right), \quad (6.8)$$

where there is only a finite number of nonzero coefficients $p_\alpha \in \mathbb{R}$ and, for $\alpha = (\alpha_1, \dots, \alpha_n) \in \mathbb{N}^n$, the *monomial* x^α is a compact writing for

$$x^\alpha := x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_n^{\alpha_n}.$$

The *degree of the monomial* x^α is denoted by $|\alpha| := \sum_{i=1}^n \alpha_i$.

We show below, that when $n = 1$, problem (6.8) can be rewritten as an SDO problem. When $n > 1$, this is usually not the case, but more and more precise SDO relaxations of the problem can be obtained. SDO relaxations can also be defined for problems with polynomial constraints; for simplicity and by the need for brevity, we will not consider that possibility here; see [108, 90, 91, 6, 18, 92] for more complete presentations.

The optimal value of problem (6.8) can be obtained by solving

$$\begin{cases} \sup_{s \in \mathbb{R}} s \\ p(x) \geq s, \quad \forall x \in \mathbb{R}^n. \end{cases}$$

This problem has a single unknown $s \in \mathbb{R}$ but an infinite number of constraints, which is not a desirable feature. This infinite number of constraints “disappears” if we trivially express them in terms of membership to the cone \mathcal{P} of nonnegative polynomials:

$$\begin{cases} \sup_{s \in \mathbb{R}} s \\ p - s \in \mathcal{P}. \end{cases} \quad (6.9)$$

Normally, we should not have gained much by these trivial transformations, unless membership to \mathcal{P} can be described in a pleasant manner. Let us look at that.

A real nonnegative number $p \in \mathbb{R}$ can always be written as the square of another number $p = q^2$ (take the square root of p for q). When p is a nonnegative polynomial, it cannot be written in general as the square of another polynomial. For example $p(x) = x^2 + 1$ cannot be written $(ax + b)^2$, since one should have both $a = b = 1$ and $ab = 0$, which are not compatible conditions. It turns out, however, that for $n = 1$ a polynomial $p \in \mathcal{P}$ can be written as the sum of *two* squares of polynomials. This is a remarkable fact and we shall see below why this observation is useful.

Proposition 6.3 (nonnegative univariate polynomial) *A univariate polynomial $p \in \mathbb{R}[x]$ is nonnegative on \mathbb{R} if and only if it is of even degree, say $2m$, and reads $p = q^2 + r^2$, with $q, r \in \mathbb{R}[x]$, $\deg q = m$, and $\deg r \leq m - 1$.*

Proof. The given conditions are clearly sufficient. Let us show that they are necessary.

Being nonnegative on \mathbb{R} the polynomial is necessary of even degree, say $2m$, and the leading coefficient is positive. There is therefore no loss of generality in supposing that this leading coefficient is 1. Then the polynomial can be decomposed in m factors of the form

$$(x - a)^2 + b^2.$$

It is indeed the form of $(x - r)(x - \bar{r})$ when r and \bar{r} are complex conjugate roots $a \pm ib$. On the other hand, any real root has an even multiplicity (otherwise the polynomial would have positive and negative values around the root) and each double root is of the form above with $b = 0$.

The m factors of the form $q^2 + r^2$ are then multiplied successively, by using the following formula

$$(q_j^2 + r_j^2)(q^2 + r^2) = (q_j q + r_j r)^2 + (q_j r - r_j q)^2 =: q_{j+1}^2 + r_{j+1}^2.$$

By induction, we see that $\deg q_j = j$ and $\deg r_j \leq j - 1$. It is indeed the case for $j = 1$ since $\deg q_1 = 1$ and $\deg r_1 = 0$. Now, be r vanishing or not, $\deg q_{j+1} = \deg q_j + 1 = j + 1$. Finally, if $r = 0$, $\deg r_{j+1} = \deg r_j + 1 \leq j$ and, if $r \neq 0$, $\deg r_{j+1} \leq \max(\deg q_j, \deg r_j + 1) \leq j$. \square

A multivariate nonnegative polynomial cannot be written as the sum of two squares of polynomials, but, on a compact set, these can be approached by SOS polynomials. An *SOS polynomial* is a polynomial that can be written as a sum of squares of polynomials and the set of SOS polynomials is denoted by

$$\Sigma[x].$$

Since $\Sigma[x] \subseteq \mathcal{P}$, problem (6.9) can be approached by

$$\begin{cases} \sup_{s \in \mathbb{R}} s \\ p - s \in \Sigma[x]. \end{cases} \quad (6.10)$$

This is a relaxation of problem (6.9) in the sense that its optimal value does not exceed the one of problem (6.9).

Proposition 6.4 tells us that a multivariate polynomial $p \in \Sigma[x]$ has a nice representation, making use of a positive semi-definite symmetric matrix S . As we shall see, a consequence of this observation is that problem (6.10) is an SDO problem, so that we have obtained an SDO relaxation of the problem consisting in finding the global minimum value of a multivariate polynomial. The proposition uses the vector $v_m(x)$, whose components are the multivariate monomials of degree $\leq m$. Since there are

$$\binom{n+d-1}{d}$$

monomials of degree d (it is the number of combination with repetition of d elements among n), the dimension of the vector $v_m(x)$ is

$$N := \binom{n-1}{0} + \binom{n}{1} + \cdots + \binom{n+d-1}{d} = \binom{n+m}{n}.$$

Hence a polynomial of degree $\leq m$ can be written $s^\top v_m(x)$, for some $s \in \mathbb{R}^N$.

Proposition 6.4 (characterization of SOS polynomials) *A multivariate polynomial $p \in \mathbb{R}[x]$ of degree $\leq 2m$ is a sum of r squares of polynomials if and only if there exists a matrix $S \succeq 0$ of order N and of rank $\leq r$ such that $p(x) = v_m(x)^\top S v_m(x)$.*

Proof. [\Rightarrow] If $p \in \mathbb{R}[x]$ is of degree $\leq 2m$ and reads $\sum_{i=1}^r \sigma_i^2$, where $\sigma_i \in \mathbb{R}[x]$, the degrees $\deg \sigma_i \leq m$. Therefore, one can find vectors $s_i \in \mathbb{R}^{m+1}$ such that

$$p(x) = \sum_{i=1}^r (s_i^\top v_m(x))^2 = \sum_{i=1}^r v_m(x)^\top s_i s_i^\top v_m(x) = v_m(x)^\top S v_m(x),$$

where $S := \sum_{i=1}^r s_i s_i^\top \succeq 0$ is of rank $\leq r$.

[\Leftarrow] Conversely, suppose that $p(x) = v_m(x)^\top S v_m(x)$, with $S \succeq 0$ of rank $\leq r$. The spectral decomposition of $S = \sum_{i=1}^r s_i s_i^\top$ allows us to write

$$p(x) = \sum_{i=1}^r v_m(x)^\top s_i s_i^\top v_m(x) = \sum_{i=1}^r (s_i^\top v_m(x))^2,$$

showing that p is the sum of the squares of at most r polynomials. □

Using the above propositions and setting $t := -s$, problem 6.10 can be written

$$\begin{cases} \inf_{(S,t) \in \mathcal{S}^{m+1} \times \mathbb{R}} t \\ p(x) + t = v_m(x)^\top S v_m(x), & \forall x \in \mathbb{R}^n \\ S \succeq 0. \end{cases} \quad (6.11)$$

Knowing p , the first constraint in (6.11) is an affine constraint on the unknown $t \in \mathbb{R}$ and the unknown real coefficients of S , if we impose equality between the coefficients of the same monomials in both sides of the equality. Therefore, problem (6.11) is almost in the primal form of an SDO problem. The difference is the free variable t , which is only constrained by the affine constraint in (6.11), not by a nontrivial cone. This small difference can be dealt with by standard SDO codes.

Rank Relaxation of a QCQP

6.1.3 Existence of Solution, Optimality Conditions

Existence of Solution

In linear optimization, according to (1.63), a finite optimal value guarantees that the problem has a solution. We have seen this result as a consequence of the fact that the image $T(P)$ of a convex polyhedron P by a linear map T is a convex polyhedron, hence a closed set. This approach no longer works for an SDO problem. Although $T(\cdot) = \langle C, \cdot \rangle$ is linear, the feasible set $\mathcal{F}_P := \{X \in \mathcal{S}^n : \mathcal{A}(X) = b, X \succeq 0\}$ is not a convex polyhedron, so that $T(P)$ may not be closed. As a result, in general

$$\text{val}(P) \in \mathbb{R} \quad \Longleftrightarrow \quad \text{Sol}(P) \neq \emptyset.$$

See example 6.7-2 below. The existence of solution is then ensured by using some kind of constraint qualification.

The assumptions of the next proposition can be understood, or at least memorized, with the following scheme in mind. If $\mathcal{F}_D^s \neq \emptyset$, a kind of Slater constraint qualification (CQ-S) holds for the dual problem (D); hence, the dual solutions of (D), which are the solutions of (P), must exist and form a bounded set (recall that (CQ-S) is equivalent to (CQ-MF) or (CQ-R) for convex problems, and use propositions 2.6 and 2.22); this is point 1 of the proposition (this way of thinking is not quite correct, since it is not assumed there that the dual problem has a solution). Similarly, if $\mathcal{F}_P^s \neq \emptyset$, a kind of Slater constraint qualification (CQ-S) holds for the primal problem (P); hence, the dual solutions of (P), which are the solutions of (D), must exist and form a bounded set; this is point 2 of the proposition (this way of thinking is not quite correct, since it is not assumed there that the primal problem has a solution). Point 3 just gathers the results in points 1 and 2.

We start with a lemma giving a consequence of the strict primal or dual feasibility.

Lemma 6.5 (consequence of strict feasibility)

- 1) If $\mathcal{F}_P^s \neq \emptyset$, then any $(d, D) \in \mathbb{R}^m \times \mathcal{S}^n$ verifying $\langle b, d \rangle \geq 0$, $\mathcal{A}^*(d) + D = 0$, $D \succeq 0$ and $d \in \mathcal{R}(\mathcal{A})$ vanishes.
- 2) If $\mathcal{F}_D^s \neq \emptyset$, then any $D \in \mathcal{S}^n$ verifying $\langle C, D \rangle \leq 0$, $\mathcal{A}(D) = 0$ and $D \succeq 0$ vanishes.

Proof. 1) Suppose that $(d, D) \in \mathbb{R}^m \times \mathcal{S}^n$ verifies $\langle b, d \rangle \geq 0$, $\mathcal{A}^*(d) + D = 0$, $D \succeq 0$ and $D \in \mathcal{R}(\mathcal{A})$. Since $\mathcal{F}_P^s \neq \emptyset$, there exists an $X_0 \in \mathcal{S}_{++}^n$ such that $\mathcal{A}(X_0) = b$. Then,

$$\begin{aligned}
 0 &= \langle X_0, \mathcal{A}^*(d) + D \rangle && [\mathcal{A}^*(d) + D = 0] \\
 &= \langle b, d \rangle + \langle X_0, D \rangle && [\mathcal{A}(X_0) = b] \\
 &\geq \langle X_0, D \rangle && [\langle b, d \rangle \geq 0] \\
 &\geq 0 && [X_0 \geq 0 \text{ and } D \geq 0].
 \end{aligned}$$

Hence $\langle X_0, D \rangle = 0$. Now $X_0 > 0$, $D \geq 0$ and (6.2) imply that $D = 0$. Next, $\mathcal{A}^*(d) = 0$ and $d \in \mathcal{R}(\mathcal{A})$ imply that $d = 0$.

2) Suppose that $D \in \mathcal{S}^n$ verifies $\langle C, D \rangle \leq 0$, $\mathcal{A}(D) = 0$ and $D \geq 0$. Since $\mathcal{F}_D^s \neq \emptyset$, there exists a pair $(y_0, S_0) \in \mathbb{R}^m \times \mathcal{S}_{++}^n$ such that $\mathcal{A}^*(y_0) + S_0 = C$. Then,

$$\begin{aligned}
 0 &= \langle y_0, \mathcal{A}(D) \rangle && [\mathcal{A}(D) = 0] \\
 &= \langle C, D \rangle - \langle S_0, D \rangle && [\mathcal{A}^*(y_0) + S_0 = C] \\
 &\leq -\langle S_0, D \rangle && [\langle C, D \rangle \leq 0] \\
 &\leq 0 && [S_0 \geq 0 \text{ and } D \geq 0].
 \end{aligned}$$

Hence $\langle S_0, D \rangle = 0$. Now $S_0 > 0$, $D \geq 0$ and (6.2) imply that $D = 0$. \square

Proposition 6.6 (existence of compact sets of solutions)

- 1) $\mathcal{F}_P \times \mathcal{F}_D^s \neq \emptyset \implies \text{Sol}(P)$ is nonempty and compact.
 - 2) $\mathcal{F}_P^s \times \mathcal{F}_D \neq \emptyset \implies \text{Sol}(D) \cap (\mathcal{R}(\mathcal{A}) \times \mathcal{S}^n)$ is nonempty and compact.
 - 3) $\mathcal{F}^s \neq \emptyset \implies \text{Sol}(P)$ and $\text{Sol}(D) \cap (\mathcal{R}(\mathcal{A}) \times \mathcal{S}^n)$ are nonempty and compact.
- In all these cases, there is no duality gap: $\text{val}(D) = \text{val}(P)$.

Proof. 1) [$\text{Sol}(P)$ nonempty and compact] Problem (P) reads

$$\inf_{X \in \mathcal{S}^n} \left(\varphi(X) := \langle C, X \rangle + \mathcal{I}_{\mathcal{A}}(X) + \mathcal{I}_{\mathcal{S}_+^n}(X) \right),$$

where $\mathcal{A} := \{X \in \mathcal{S}^n : \mathcal{A}(X) = b\}$. The value at $D \in \mathcal{S}^n$ of the asymptotic function φ^∞ of φ reads

$$\varphi^\infty(D) = \langle C, D \rangle + \mathcal{I}_{\mathcal{N}(\mathcal{A})}(D) + \mathcal{I}_{\mathcal{S}_+^n}(D).$$

By the implication (iv) \implies (iii) of proposition 1.20, $\text{Sol}(P)$ is nonempty and compact if we have that $\varphi^\infty(D) > 0$ for all nonzero D or, equivalently, if

$$\mathcal{A}(D) = 0, \quad D \geq 0, \quad \langle C, D \rangle \leq 0 \quad \implies \quad D = 0. \quad (6.12a)$$

By lemma 6.5, this is a consequence of $\mathcal{F}_D^s \neq \emptyset$.

[No duality gap] Consider first the case when $C \in \mathcal{R}(\mathcal{A}^*) = \mathcal{N}(\mathcal{A})^\perp$ (see figure 6.1, left). Then, C is perpendicular to $\mathcal{N}(\mathcal{A})$, any primal feasible $X \in \mathcal{F}_P$ should be a solution and the optimal value $\text{val}(P)$ should not be affected by a perturbation of \mathcal{S}_+^n , so that $S_0 = 0$ should be an optimal dual variable. Let us check this rigorously. Indeed, since when $C \in \mathcal{R}(\mathcal{A}^*)$, one can find a $y_0 \in \mathbb{R}^m$ such that $C = \mathcal{A}^*(y_0)$, we have for an arbitrary $X \in \mathcal{F}_P$: $\text{val}(P) = \langle C, X \rangle = \langle y_0, \mathcal{A}(X) \rangle = \langle b, y_0 \rangle$. But $(y_0, 0) \in \mathcal{F}_D$, hence $\text{val}(D) \geq \langle b, y_0 \rangle$, which shows the absence of duality gap.

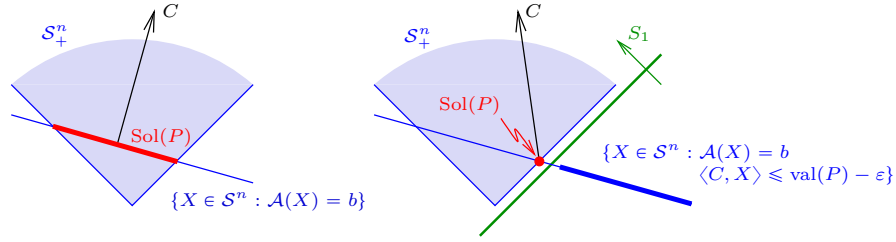


Fig. 6.1. Artistic 2D illustration of the proof of proposition 6.6 (\mathcal{S}_+^n is much more complex than in the given representation): the simple case when $C \in \mathcal{N}(\mathcal{A})^\perp$ (left) and the separation technique when $C \notin \mathcal{N}(\mathcal{A})^\perp$ (right)

Consider now the case when $C \notin \mathcal{R}(\mathcal{A}^*) = \mathcal{N}(\mathcal{A})^\perp$ (see figure 6.1, right). Take $\varepsilon > 0$. It suffices to find a $(y_0, S_0) \in \mathcal{F}_D$ such that $\langle b, y_0 \rangle \geq \text{val}(P) - \varepsilon$. The dual variable S_0 will be obtained by a separation argument, while y_0 will result as a by-product. Consider the two sets

$$C_1 := \mathcal{S}_+^n \quad \text{and} \quad C_2 := \{X \in \mathcal{S}^n : \mathcal{A}(X) = b, \langle C, X \rangle \leq \text{val}(P) - \varepsilon\}.$$

These sets

- are closed and convex (clear),
- are nonempty (for C_2 , take $H \in \mathcal{N}(\mathcal{A})$ such that $\langle C, H \rangle < 0$ [possible since $C \notin \mathcal{R}(\mathcal{A}^*)$] and observe that $\mathcal{A}(X + tH) = b$ and $\langle C, X + tH \rangle \rightarrow -\infty$ when $t \rightarrow \infty$),
- are disjoint (by the definition of $\text{val}(P)$) and
- verify $C_1^\infty \cap C_2^\infty = \{0\}$ (indeed, if $D \in C_1^\infty = \mathcal{S}_+^n$ and $D \in C_2^\infty = \{D \in \mathcal{S}^n : \mathcal{A}(D) = 0, \langle C, D \rangle \leq 0\}$, we have $D = 0$ by (6.12a)).

Therefore, C_1 and C_2 can be strictly separated (point 2 of proposition 1.15): there exists $S_1 \in \mathcal{S}^n$ such that

$$\alpha := \sup_{\substack{\mathcal{A}(X)=b \\ \langle C, X \rangle \leq \text{val}(P) - \varepsilon}} \langle S_1, X \rangle < \inf_{X \geq 0} \langle S_1, X \rangle =: \beta.$$

Take in the right-hand side $X = 0$ to get $\alpha < 0$ and $X = tvv^\top$ with $t \rightarrow \infty$ to get $S_1 \geq 0$. Next, observe that the problem in the left-hand side is a linear optimization problem on the vector space \mathcal{S}^n , whose optimal value is finite. By (1.63), there exist multipliers $(y_1, \sigma_1) \in \mathbb{R}^m \times \mathbb{R}$ such that

$$\mathcal{A}^*(y_1) + S_1 = \sigma_1 C, \quad \sigma_1 \geq 0 \quad \text{and} \quad -\langle b, y_1 \rangle + \sigma_1(\text{val}(P) - \varepsilon) = \alpha < 0.$$

We necessarily have $\sigma_1 > 0$, since otherwise, the last inequality would yield $\langle b, y_1 \rangle > 0$ and, scalarly multiplying the first identity by an $X \in \mathcal{F}_P$ would give $\langle b, y_1 \rangle = -\langle X, S_1 \rangle \leq 0$, a contradiction. Now setting $y_0 = y_1/\sigma_1$ and $S_0 := S_1/\sigma_1$, we get $(y_0, S_0) \in \mathcal{F}_D$ verifying $\langle b, y_0 \rangle > \text{val}(P) - \varepsilon$, as desired.

2) The claim can be shown like in the proof of point 1 (exercise 6.1.2).

3) This is a consequence of points 1 and 2. □

The result in point 2 and 3 of proposition 6.6 simplify if \mathcal{A} is surjective, since then they become

$\mathcal{F}_P^s \times \mathcal{F}_D \neq \emptyset, \mathcal{A}$ surjective $\implies \text{Sol}(D)$ is nonempty and compact.
 $\mathcal{F}^s \neq \emptyset, \mathcal{A}$ surjective $\implies \text{Sol}(P) \times \text{Sol}(D)$ is nonempty and compact.

If \mathcal{A} is not surjective, either $b \notin \mathcal{R}(\mathcal{A})$, in which case the feasible set $\mathcal{F}_P = \emptyset$ and $\text{val}(P) = +\infty$, or $b \in \mathcal{R}(\mathcal{A})$. In the latter case, when \mathcal{A} has the representation (6.6a), one can remove redundant constraints $\langle A_k, X \rangle = b_k$ and end up with a linear independent set of matrices A_k forming a new surjective operator \mathcal{A} . Therefore, numerically, adding an assumption on the surjectivity of \mathcal{A} is not dramatic but, in theory, it is more elegant and it can be useful to avoid it.

We conclude this subsection by giving some examples and counter-examples related to proposition 6.6.

Examples 6.7 Here are some situations that can occur (the proof of the claims made below are proposed in the exercise 6.1.3).

1) \mathcal{A} is surjective, (P) has a unique solution, $\mathcal{F}_D^s \neq \emptyset, \text{val}(D) \in \mathbb{R}$ but (D) has no solution, there is no duality gap.

This situation is compatible with point 1 of proposition 6.6. Here is an example with $n = 2$ and $m = 2$:

$$C = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_2 = \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix}, \quad b = \begin{pmatrix} -1 \\ 0 \end{pmatrix}.$$

2) \mathcal{A} is surjective, $\mathcal{F}_P^s \neq \emptyset, \text{val}(P) \in \mathbb{R}$ but (P) has no solution, (D) has a unique solution but $\mathcal{F}_D^s = \emptyset$, there is no duality gap.

This situation is compatible with point 2 of proposition 6.6. Here is an example with $n = 2$ and $m = 1$:

$$C = \begin{pmatrix} 2 & -1 \\ -1 & 0 \end{pmatrix}, \quad A_1 = \begin{pmatrix} 0 & -1 \\ -1 & 2 \end{pmatrix} \quad \text{and} \quad b = 2. \quad \square$$

Optimality Conditions

Proposition 6.8 (optimality conditions) Suppose that $\mathcal{F}_P^s \neq \emptyset$ or $\mathcal{F}_D^s \neq \emptyset$, then

$$(X, (y, S)) \in \text{Sol}(P) \times \text{Sol}(D) \iff \begin{cases} \mathcal{A}^*(y) + S = C, & S \geq 0, \\ \mathcal{A}(X) = b, & X \geq 0, \\ \langle X, S \rangle = 0. \end{cases} \quad (6.13)$$

Proof. [\implies] The first two conditions are the dual and primal feasibility, which is satisfied by a primal-dual solution. The third condition is the absence of duality gap, guaranteed by proposition 6.6.

[\impliedby] This is point 3 of proposition 6.2. □

6.1.4 Interior Point Algorithms \blacktriangle

Central Path

There are good reasons to generate iterates well inside \mathcal{F}_P^s . This is obtained analytically (not geometrically) by an [interior penalization](#):

$$\begin{aligned} (P) \quad &\curvearrowright \quad (P_\mu) \quad \begin{cases} \inf_X \langle C, X \rangle + \mu \text{ld}(X) \\ \mathcal{A}(X) = b, \end{cases} \\ (D) \quad &\curvearrowright \quad (D_\mu) \quad \begin{cases} \sup_{(y,S)} \langle b, y \rangle - \mu \text{ld}(S) \\ \mathcal{A}^*(y) + S = C, \end{cases} \end{aligned}$$

where $\text{ld} : \mathcal{S}^n \rightarrow \overline{\mathbb{R}}$ is the [strictly convex](#) and [closed](#) function defined at X by

$$\text{ld}(X) := \begin{cases} -\log \det(X) & \text{if } X > 0 \\ +\infty & \text{otherwise.} \end{cases}$$

Here are three properties of ld (with $X > 0$ and $H, K \in \mathcal{S}^n$):

$$\begin{aligned} \text{ld}'(X) \cdot H &= -\langle X^{-1}, H \rangle, \\ \text{ld}''(X) \cdot (H, K) &= \langle X^{-1} H X^{-1}, K \rangle, \\ \text{ld}^\infty &= \mathcal{I}_{\mathcal{S}_+^n}. \end{aligned}$$

The [central path](#) is the smooth curve $\mathcal{C} : \mu \in \mathbb{R}_{++} \mapsto$ the unique solution to

$$(O_\mu) \quad \begin{cases} \mathcal{A}^*(y) + S = C, & S > 0, \\ \mathcal{A}(X) = b, & X > 0, \\ XS = \mu I. \end{cases} \quad (6.14)$$

Lemma 6.9 (equivalent definitions of the central path) *Suppose that $\mathcal{F}^s \neq \emptyset$ and $\mu > 0$. Then,*

- 1) $X_\mu \in \mathcal{S}_{++}^n$ solves problem (P_μ) if and only if there is a pair $(y_\mu, S_\mu) \in \mathbb{R}^m \times \mathcal{S}_{++}^n$ such that (O_μ) holds,
- 2) $(y_\mu, S_\mu) \in \mathbb{R}^m \times \mathcal{S}_{++}^n$ solves problem (D_μ) if and only if there is an $X_\mu \in \mathcal{S}_{++}^n$ such that (O_μ) holds.

Proof. We alleviate the notation by dropping the subscript μ on the variables.

1) By the convexity of the objective of problem (P_μ) (ld is a convex function) and the affinity of its constraint (hence it is qualified by [\(CQ-A\)](#)), $X \in \mathcal{S}_{++}^n$ solves (P_μ) if and only if it verifies $\mathcal{A}(X) = b$ and there exists a multiplier $y \in \mathbb{R}^m$ such that the gradient of the Lagrangian of the problem vanishes:

$$C - \mu X^{-1} - \mathcal{A}^*(y) = 0.$$

Setting $S := \mu X^{-1} > 0$, we get (O_μ) .

2) By the concavity of the objective of problem (D_μ) and the affinity of its constraint, $(y, S) \in \mathbb{R}^m \times \mathcal{S}_{++}^n$ solves (D_μ) if and only if it verifies $\mathcal{A}^*(y) + S = C$ and

there exists a multiplier $X \in \mathcal{S}^n$ such that the gradient of the Lagrangian of the problem vanishes:

$$-b + \mathcal{A}(X) = 0 \quad \text{and} \quad -\mu S^{-1} + X = 0.$$

Hence $X = \mu S^{-1} > 0$ and we get also (O_μ) . □

Proposition 6.10 (existence and smoothness of the central path) *Suppose that $\mathcal{F}^s \neq \emptyset$ and $\mu > 0$. Then,*

- 1) *the system (O_μ) has a solution (X_μ, y_μ, S_μ) , unique in $\mathcal{S}_{++}^n \times \mathcal{R}(\mathcal{A}) \times \mathcal{S}_{++}^n$,*
- 2) *the map $\mu \in \mathbb{R}_{++} \mapsto (X_\mu, y_\mu, S_\mu) \in \mathcal{S}_{++}^n \times \mathcal{R}(\mathcal{A}) \times \mathcal{S}_{++}^n$ is C^∞ .*

Proof. We alleviate the notation by dropping the subscript μ on the variables.

1) By lemma 6.9, (X, y, S) solves (O_μ) , if and only if X solves (P_μ) and (y, S) is a pair of optimal multipliers associated with the constraints of that problem.

Therefore, to prove the existence of a solution to (O_μ) , it suffices to show that (P_μ) has a solution. Now, problem (P_μ) also reads

$$\inf_{X \in \mathcal{S}^n} \left(\varphi_\mu(X) := \langle C, X \rangle + \mu \text{ld}(X) + \mathcal{I}_{\mathcal{A}}(X) \right),$$

where $\mathcal{A} := \{X \in \mathcal{S}^n : \mathcal{A}(X) = b\}$. The value at $D \in \mathcal{S}^n$ of the asymptotic function φ_μ^∞ of φ_μ reads

$$\varphi_\mu^\infty(D) = \langle C, D \rangle + \mathcal{I}_{\mathcal{S}_+^n}(D) + \mathcal{I}_{\mathcal{N}(\mathcal{A})}(D).$$

Like in the proof of proposition 6.6, $\mathcal{F}_D^s \neq \emptyset$ implies that $\varphi_\mu^\infty(D) > 0$ for $D \neq 0$. By the implication (iv) \Rightarrow (iii) of proposition 1.20, φ_μ has a nonempty and compact set of minimizers.

The uniqueness of the solution X_μ of (O_μ) comes from the strict convexity of ld , which ensures the uniqueness in X_μ of the solution to (P_μ) and the recalled equivalence between the solutions to (P_μ) and (O_μ) . The uniqueness of S_μ comes from the uniqueness of X_μ and the equation $X_\mu S_\mu = \mu I$, which determines S_μ from X_μ . The uniqueness of y_μ in $\mathcal{R}(\mathcal{A})$ comes from the following observations. Since y only intervenes in (O_μ) through the equation

$$\mathcal{A}^*(y) + S = C,$$

and since any solution y_μ can be written $y_\mu = y_\mu^0 + y_\mu^1$ with $y_\mu^0 \in \mathcal{N}(\mathcal{A}^*)$ and $y_\mu^1 \in \mathcal{N}(\mathcal{A}^*)^\perp = \mathcal{R}(\mathcal{A})$, (X_μ, y_μ^1, S_μ) is also a solution to (O_μ) with its y -component in $\mathcal{R}(\mathcal{A})$. Now, if (X_μ, y_μ, S_μ) and (X_μ, y'_μ, S_μ) are two solutions to (O_μ) with a y -component in $\mathcal{R}(\mathcal{A})$, $h := y'_\mu - y_\mu$ satisfies $h \in \mathcal{R}(\mathcal{A})$ and $h \in \mathcal{N}(\mathcal{A}^*) = \mathcal{R}(\mathcal{A})^\perp$, hence $h = 0$ and $y_\mu = y'_\mu$.

2) We have just shown, actually, that, when $\mathcal{F}^s \neq \emptyset$ and $\mu > 0$, (X_μ, y_μ, S_μ) is the unique solution to the system

$$\begin{cases} \mathcal{A}^*(y) + S = C, \\ \mathcal{A}(X) = b, \\ XS = \mu I \end{cases}$$

in the open set $\mathcal{S}_{++}^n \times \mathcal{R}(\mathcal{A}) \times \mathcal{S}_{++}^n$ of the vector space $\mathcal{S}^n \times \mathcal{R}(\mathcal{A}) \times \mathcal{S}^n$. Thanks to this system, the smoothness of the map $\mu \in (0, +\infty) \mapsto z_\mu := (X_\mu, y_\mu, S_\mu)$ can be derived from the implicit function theorem. Indeed, (z_μ, μ) is a zero of $F(z, \mu)$, where

$$F : (\mathcal{S}^n \times \mathcal{R}(\mathcal{A}) \times \mathcal{S}^n) \times \mathbb{R} \rightarrow \mathcal{S}^n \times \mathcal{R}(\mathcal{A}) \times \mathcal{S}^n$$

is defined at $(z, \mu) \in (\mathcal{S}^n \times \mathcal{R}(\mathcal{A}) \times \mathcal{S}^n) \times \mathbb{R}$ by

$$F(z, \mu) = \begin{pmatrix} \mathcal{A}^*(y) + S - C \\ \mathcal{A}(X) - b \\ XS - \mu I \end{pmatrix}.$$

Since F is infinitely continuously differentiable, this smoothness property will be inherited by the implicit function $\mu \mapsto z_\mu$, provide $F'_z(z_\mu, \mu)$ is bijective, or even injective since for fixed μ the dimension of the range space of F is equal to the one of its definition space. Therefore, we only have to show that

$$\begin{pmatrix} 0 & \mathcal{A}^* & I \\ \mathcal{A} & 0 & 0 \\ S & 0 & X \end{pmatrix} \begin{pmatrix} d_X \\ d_y \\ d_S \end{pmatrix} = 0 \quad \text{and} \quad d_y \in \mathcal{R}(\mathcal{A})$$

implies that $(d_X, d_y, d_S) = 0$. From the last row, we get $d_X = -S^{-1}X d_S$, which with the second row yields $\mathcal{A}S^{-1}X d_S = 0$. Since $d_S = -\mathcal{A}^*d_y$ by the first row, we get $\mathcal{A}S^{-1}X\mathcal{A}^*d_y = 0$. Now $S^{-1}X = \mu S^{-2} > 0$, so that $\mathcal{A}^*d_y = 0$ and finally $d_y = 0$ by the fact that $d_y \in \mathcal{R}(\mathcal{A})$. Next $d_S = -\mathcal{A}^*d_y = 0$ and $d_X = -S^{-1}X d_S = 0$. \square

A Convergent Algorithm With Feasible Iterates

A primal-dual path-following interior-point algorithm generates iterates

$$z_k := (X_k, y_k, S_k) \in \mathcal{F}^s$$

in a neighborhood $V(\theta)$ of the central path \mathcal{C} ($\theta \in (0, 1)$ is a parameter that determines the size of the neighborhood). Each iteration proceeds along a Newton direction aiming a moving point on \mathcal{C} , whose central parameter is $\sigma\bar{\mu}(z)$ where $\sigma \in (0, 1)$ and

$$\bar{\mu}(z) := \frac{\langle X, S \rangle}{n}.$$

Algorithm 6.11 (interior point scheme in SDO) One iteration, from $z := (X, y, S)$ to the next one $z_+ := (X_+, y_+, S_+)$.

1. *Newton step:* Let d be the Newton direction on a symmetrized version of $(O_{\sigma\bar{\mu}(z)})$.
2. *Displacement:* Determine a large stepsize $\alpha > 0$ such that $z + \alpha d \in V(\theta)$.
3. *New iterate:* $z_+ := z + \alpha d$.

A Practical Algorithm With Infeasible Iterates**6.1.5 A Nonsmooth Algorithm ▲****Notes**

This chapter owes a lot to the reviews of Alizadeh [3; 1995], Vandenberghe and Boyd [135; 1996], Monteiro and Todd [102; 2000], Todd [132; 2001] and to the book of Nemirovskii and Nesterov [105; 1994]. The monographs of Saigal, Vandenberghe and Wolkowicz [136; 2000], de Klerk [42; 2002] and Anjos and Lasserre [5; 2012] are rich in information on the theory, algorithms and applications. Monteiro [101; 2003] reviews methods for solving SDO problems, including approaches that do not use the notion of interior points.

The proof of proposition 6.3 is taken from [111; 1976; VI 44]. The presented proof of the absence of duality gap in proposition 6.6 is due to Todd [132; theorem 4.1].

Exercises

6.1.1. *Schur complement of a positive semi-definite matrix.* Let $A \in \mathcal{S}^n$, $C \in \mathcal{S}^m$, and $B \in \mathbb{R}^{n \times m}$. Denote by A^\dagger the pseudo-inverse of A . Show that

$$\begin{pmatrix} A & B \\ B^\top & C \end{pmatrix} \succeq 0 \iff \begin{cases} A \succeq 0 \\ \mathcal{R}(B) \subseteq \mathcal{R}(A) \\ C - B^\top A^\dagger B \succeq 0. \end{cases} \quad (6.15)$$

6.1.2. *Existence of a nonempty compact solution set and no duality gap.* Prove point 2 of proposition 6.6 and the absence of duality gap in that case.

6.1.3. *SDO examples.* Prove the claims made in examples 6.7.

6.2 Circular Optimization ▲**6.3 Copositive Optimization ▲**

Appendices

References

- [1] V. Acary, B. Brogliato (2008). *Numerical Methods for Nonsmooth Dynamical Systems - Applications in Mechanics and Electronics*. Lecture Notes in Applied and Computational Mechanics 35. Springer. [\[doi\]](#). 129
- [2] S. Adly, R.T. Rockafellar (2019). Sensitivity analysis of monotone inclusions via the proto-differentiability of the resolvent operator. Technical report. 98
- [3] F. Alizadeh (1995). Interior point methods in semidefinite programming with applications to combinatorial optimization. *SIAM Journal on Optimization*, 5, 13–51. 170
- [4] L. Ambrosio, P. Tilli (2004). *Topics on Analysis in Metric Spaces*. Oxford Lecture Series in Mathematics and its Applications 25. Oxford University Press, Oxford. 31
- [5] M.F. Anjos, J.B. Lasserre (2012). *Handbook on Semidefinite, Conic and Polynomial Optimization*, volume 166 of *International Series in Operations Research & Management Science*. Springer. [\[doi\]](#). 170
- [6] M.F. Anjos, J.B. Lasserre (2012). *Introduction to semidefinite, conic and polynomial optimization*, volume 166 of *International Series in Operations Research & Management Science*, chapter 1. Springer. 160
- [7] A. Auslender, M. Teboulle (2003). *Asymptotic Cones and Functions in Optimization and Variational Inequalities*. Springer Monographs in Mathematics. Springer, New York. 26
- [8] M. Avriel (1976). *Nonlinear Programming, Analysis and Methods*. Prentice-Hall, Englewood Cliffs, New Jersey. 90
- [9] L. Beaudé, K. Brenner, S. Lopez, R. Masson, F. Smay (2019). Non-isothermal compositional liquid gas Darcy flow: formulation, soil-atmosphere boundary condition and application to high-energy geothermal simulations. *Computational Geosciences*, 23(3), 443–470. [\[doi\]](#). 129
- [10] I. Ben Gharbia, J. Dabaghi, V. Martin, M. Vohralík (2020). A posteriori error estimates for a compositional two-phase flow with nonlinear complementarity constraints. *Computational Geosciences*, 24(3), 1031–1055. [\[doi\]](#). 129
- [11] I. Ben Gharbia, E. Flauraud (2019). Study of compositional multiphase flow formulation using complementarity conditions. *Oil & Gas Sciences and Technology*, 74, 1–15. [\[doi\]](#). 129
- [12] I. Ben Gharbia, J. Jaffré (2014). Gas phase appearance and disappearance as a problem with complementarity constraints. *Mathematics and Computers in Simulation*, 99, 28–36. [\[doi\]](#). 129
- [13] A. Ben-Tal (1980). Second-order and related extremality conditions in nonlinear programming. *Journal of Optimization Theory and Applications*, 31(2), 143–165. [\[doi\]](#). 92
- [14] A. Ben-Tal, A. Nemirovski (2001). *Lectures on Modern Convex Optimization – Analysis, Algorithms, and Engineering Applications*. MPS-SIAM Series on Optimization 2. SIAM. 155
- [15] A. Ben-Tal, J. Zowe (1982). A unified theory of first and second order conditions for extremum problems in topological vector spaces. *Mathematical Programming Study*, 19, 39–76. [\[doi\]](#). 92

- [16] A. Berman, N. Shaked-Monderer (2003). *Completely Positive Matrices*. World Scientific, River Edge, NJ, USA. [90](#)
- [17] D.P. Bertsekas (1982). Projected Newton methods for optimization problems with simple constraints. *SIAM Journal on Control and Optimization*, 20, 221–246. [\[doi\]](#). [152](#)
- [18] G. Blekherman, P.A. Parrilo, R.R. Thomas (2013). *Semidefinite Optimization and Convex Algebraic Geometry*. MOS-SIAM Series on Optimization. SIAM and MPS, Philadelphia. [\[doi\]](#). [160](#)
- [19] J.F. Bonnans (1994). Local analysis of Newton-type methods for variational inequalities and nonlinear programming. *Applied Mathematics and Optimization*, 29, 161–186. [\[doi\]](#). [102](#), [112](#), [133](#), [152](#)
- [20] J.F. Bonnans, J.Ch. Gilbert, C. Lemaréchal, C. Sagastizábal (2006). *Numerical Optimization – Theoretical and Practical Aspects* (second edition). Universitext. Springer Verlag, Berlin. [\[authors\]](#) [\[editor\]](#) [\[doi\]](#). [152](#)
- [21] J.F. Bonnans, A. Shapiro (2000). *Perturbation Analysis of Optimization Problems*. Springer Verlag, New York. [iii](#)
- [22] J.M. Borwein, A.S. Lewis (2000). *Convex Analysis and Nonlinear Optimization – Theory and Examples*. CMS Books in Mathematics 3. Springer, New York. [26](#)
- [23] J.M. Borwein, Q.J. Zhu (2010). *Techniques of Variational Analysis*. Computational Management Science. Springer Science+Business Media, Inc., Berlin. [iii](#)
- [24] H. Brézis (1983). *Analyse Fonctionnelle Appliquée*. Masson, Paris. [62](#), [80](#)
- [25] B. Brogliato (2016). *Nonsmooth Mechanics - Models, Dynamics and Control* (third edition). Springer. [\[doi\]](#). [129](#)
- [26] H. Buchholzer, Ch. Kanzow, P. Knabner, S. Kräutle (2011). The semismooth Newton method for the solution of reactive transport problems including mineral precipitation-dissolution reactions. *Computational Optimization and Applications*, 50(2), 193–221. [\[doi\]](#). [129](#)
- [27] Q.M. Bui, H.C. Elman (2020). Semi-smooth Newton methods for nonlinear complementarity formulation of compositional two-phase flow in porous media. *Journal of Computational Physics*, 407, 109163. [\[doi\]](#). [129](#)
- [28] J.V. Burke (1991). An exact penalization viewpoint of constrained optimization. *SIAM Journal on Control and Optimization*, 29, 968–998. [92](#)
- [29] C. Carathéodory (1907). Über den Variabilitätsbereich der Koeffizienten von Potenzreihen, die gegebene Werte nicht annehmen. *Mathematische Annalen*, 64, 95–115. [12](#)
- [30] Jein-Shan Chen (2006). The semismooth-related properties of a merit function and a descent method for the nonlinear complementarity problem. *Journal of Global Optimization*, 36(4), 565–580. [\[doi\]](#). [130](#)
- [31] S.J. Chung (1989). NP-completeness of the linear complementarity problem. *Journal of Optimization Theory and Applications*, 60, 393–399. [\[doi\]](#). [129](#)
- [32] F.H. Clarke (1983). *Optimization and Nonsmooth Analysis*. John Wiley & Sons, New York. Reprinted in 1990 by SIAM, Classics in Applied Mathematics 5 [\[doi\]](#). [31](#), [119](#), [124](#)
- [33] F.H. Clarke (1990). *Optimization and Nonsmooth Analysis* (second edition). Classics in Applied Mathematics 5. SIAM, Philadelphia, PA, USA. [\[doi\]](#). [31](#)
- [34] L. Collatz, W. Wetterling (1975). *Optimization Problems*. Springer-Verlag, New York. [90](#)
- [35] R. Cominetti (1990). Metric regularity, tangent sets, and second-order optimality conditions. *Applied Mathematics and Optimization*, 21(1), 265–287. [\[doi\]](#). [66](#)
- [36] R.W. Cottle (1964). *Nonlinear Programs with Positively Bounded Jacobians*. PhD Thesis, University of California, Berkeley, USA. [129](#)

- [37] R.W. Cottle (1966). Nonlinear programs with positively bounded jacobians. *SIAM Journal on Applied Mathematics*, 14, 147–158. [\[doi\]](#). 129
- [38] R.W. Cottle (1974). Manifestations of the Schur complement. *Linear Algebra and its Applications*, 8, 189–211. 160
- [39] R.W. Cottle (2005). Linear complementarity since 1978. In *Variational analysis and applications*, Nonconvex Optimization and Its Applications 79, pages 239–257. Springer, New York. [\[doi\]](#). 128
- [40] R.W. Cottle, J.-S. Pang, R.E. Stone (2009). *The Linear Complementarity Problem*. Classics in Applied Mathematics 60. SIAM, Philadelphia, PA, USA. [\[doi\]](#). 128
- [41] J. Dabaghi, V. Martin, M. Vohralík (2017). Adaptive inexact semismooth Newton methods for the contact problem between two membranes. Technical report, INRIA. [\[hal-01666845\]](#). 129
- [42] E. de Klerk (2002). *Aspects of Semidefinite Programming - Interior Point Algorithms and Selected Applications*. Kluwer Academic Publishers, Dordrecht. [\[doi\]](#). 170
- [43] P.J.C. Dickinson, L. Gijben (2011). On the computational complexity of membership problems for the completely positive cone and its dual. Technical report. [\[Optimization Online\]](#). 90
- [44] A.L. Dontchev, R.T. Rockafellar (2009). *Implicit Functions and Solution Mappings – A View from Variational Analysis*. Springer Monographs in Mathematics. Springer. iii, 80
- [45] A.L. Dontchev, R.T. Rockafellar (2013). Convergence of inexact Newton methods for generalized equations. *Mathematical Programming*, 139, 115–137. [\[doi\]](#). 112
- [46] A.L. Dontchev, R.T. Rockafellar (2014). *Implicit Functions and Solution Mappings – A View from Variational Analysis* (second edition). Springer Series in Operations Research and Financial Engineering. Springer. 98, 112
- [47] B.C. Eaves (1971). On the basic theorem of complementarity. *Mathematical Programming*, 1, 68–75. 115
- [48] K. Erleben (2013). Numerical methods for linear complementarity problems in physics-based animation. In *ACM SIGGRAPH 2013 Courses*, SIGGRAPH '13, pages 8:1–8:42. ACM, New York, NY, USA. [\[doi\]](#). 129
- [49] L.C. Evans, R.F. Gariepy (2015). *Measure Theory and Fine Properties of Functions* (revised edition). CRC Press, Boca Raton. 29, 31
- [50] F. Facchinei, J.-S. Pang (2003). *Finite-Dimensional Variational Inequalities and Complementarity Problems* (two volumes). Springer Series in Operations Research. Springer. iii, 112, 124, 128, 129
- [51] F. Facchinei, J. Soares (1997). A new merit function for nonlinear complementarity problems and a related algorithm. *SIAM Journal on Optimization*, 7(1), 225–247. [\[doi\]](#). 130
- [52] M.C. Ferris, J.-S. Pang (1997). Engineering and economic applications of complementarity problems. *SIAM Review*, 39, 669–713. [\[doi\]](#). 129
- [53] A.V. Fiacco, G.P. McCormick (1968). *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. John Wiley, New York. Reprinted in 1990 by SIAM in the collection “Classics in Applied Mathematics”, number 4, [\[doi\]](#). 90, 92
- [54] A. Fischer (1992). A special Newton-type optimization method. *Optimization*, 24, 269–284. [\[doi\]](#). 130
- [55] R. Fletcher (2010). The sequential quadratic programming method. In G. Di Pillo, F. Schoen (editors), *Numerical Optimization*, Lecture Notes in Mathematics 1989, pages 165–214. Springer. 152
- [56] M. Frank, P. Wolfe (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3, 95–110. [\[doi\]](#). 138, 147
- [57] M. Fréchet (1911). Sur la notion de différentielle. *C. R. Acad. Sci. Paris*, 152, 845–847. [\[Gallica\]](#). 28

- [58] Y. Gao, H. Song, X. Wang, K. Zhang (2020). Primal-dual active set method for pricing American better-of option on two assets. *Communications in Nonlinear Science and Numerical Simulation*, 80. [\[doi\]](#). 129
- [59] J. Gauvin (1977). A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming. *Mathematical Programming*, 12, 136–138. [40](#), [80](#)
- [60] J. Gauvin (1992). *Théorie de la programmation mathématique non convexe*. Les Publications CRM, Montréal. [92](#)
- [61] J.Ch. Gilbert, J. Nocedal (1992). Global convergence properties of conjugate gradient methods for optimization. *SIAM Journal on Optimization*, 2, 21–42. [\[doi\]](#). [48](#)
- [62] C. Gonzaga (2000). Two facts on the convergence of the Cauchy algorithm. *Journal of Optimization Theory and Applications*, 107, 591–600. [\[doi\]](#). [49](#)
- [63] P.T. Harker, J.-S. Pang (1990). Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications. *Mathematical Programming*, 48, 161–220. [\[doi\]](#). [112](#), [129](#)
- [64] P. Hartman, G. Stampacchia (1966). On some non-linear elliptic differential-functional equations. *Acta Mathematica*, 115, 271–310. [\[doi\]](#). [112](#)
- [65] J.-B. Hiriart-Urruty, C. Lemaréchal (1993). *Convex Analysis and Minimization Algorithms*. Grundlehren der mathematischen Wissenschaften 305-306. Springer. [124](#)
- [66] J.-B. Hiriart-Urruty, C. Lemaréchal (1996). *Convex Analysis and Minimization Algorithms* (second edition). Grundlehren der mathematischen Wissenschaften 305-306. Springer. [26](#)
- [67] J.-B. Hiriart-Urruty, C. Lemaréchal (2001). *Fundamentals of Convex Analysis*. Springer. [26](#)
- [68] J.-B. Hiriart-Urruty, A. Seeger (2010). A variational approach to copositive matrices. *SIAM Review*, 52, 593–629. [90](#)
- [69] R.B. Holmes (1973). Smoothness of certain metric projections on Hilbert space. *Translations of the American Mathematical Society*, 184, 87–100. [20](#)
- [70] Kh.D. Ikramov, N.V. Savel'eva (2000). Conditionally definite matrices. *Journal of Mathematical Sciences*, 98, 1–50. [90](#)
- [71] A. Ioffe (1989). On some recent developments in the theory of second order optimality conditions. In S. Dolecki (editor), *Optimization*, Lecture Notes in Mathematics 1405, pages 54–68. Springer, Berlin. [53](#)
- [72] A.D. Ioffe (1979). Necessary and sufficient conditions for a local minimum. 3: second order conditions and augmented duality. *SIAM Journal on Control and Optimization*, 17, 266–288. [\[doi\]](#). [92](#)
- [73] G. Isac (1992). *Complementarity Problems*. Lecture Notes in Mathematics 1528. Springer, Berlin. [\[doi\]](#). [128](#), [129](#)
- [74] K. Ito, K. Kunisch (2008). *Lagrange Multiplier Approach to Variational Problems and Applications*. Advances in Design and Control. SIAM Publication, Philadelphia. [\[doi\]](#). [128](#)
- [75] A.F. Izmailov, M.V. Solodov (2010). Inexact Josephy-Newton framework for generalized equations and its applications to local analysis of Newtonian methods for constrained optimization. *Computational Optimization and Applications*, 46(2), 347–368. [\[doi\]](#). [112](#)
- [76] A.F. Izmailov, M.V. Solodov (2014). *Newton-Type Methods for Optimization and Variational Problems*. Springer Series in Operations Research and Financial Engineering. Springer. [\[doi\]](#). [iii](#), [112](#), [128](#), [129](#)
- [77] H. Jiang, L. Qi (1997). A new nonsmooth equations approach to nonlinear complementarity problems. *SIAM Journal on Control and Optimization*, 35, 178–193. [\[doi\]](#). [131](#)

- [78] H. Jiang, D. Ralph (1999). Global and local superlinear convergence analysis of Newton-type methods for semismooth equations with smooth least squares. In M. Fukushima, L. Qi (editors), *Reformulation: nonsmooth, piecewise smooth, semismooth and smoothing methods*, Applied Optimization 22, pages 181–209. Springer-Science+Business Media, B.V. [127](#)
- [79] N.H. Josephy (1979). Newton’s method for generalized equations. Technical Summary Report 1965, Mathematics Research Center, University of Wisconsin, Madison, WI, USA. [112](#)
- [80] N.H. Josephy (1979). Quasi-Newton’s method for generalized equations. Summary Report 1966, Mathematics Research Center, University of Wisconsin, Madison, WI, USA. [112](#)
- [81] S. Karamardian (1971). Generalized complementarity problem. *Journal of Optimization Theory and Applications*, 8(3), 161–168. [\[doi\]. 112](#)
- [82] W.E. Karush (1939). *Minima of Functions of Several Variables with Inequalities as Side Conditions*. Master’s thesis, Department of Mathematics, University of Chicago, Chicago. [39](#)
- [83] D. Klatte, B. Kummer (2002). *Nonsmooth Equations in Optimization - Regularity, Calculus, Methods and Applications*, volume 60 of *Nonconvex Optimization and Its Applications*. Kluwer Academic Publishers, Dordrecht. [\[doi\]. iii, 128](#)
- [84] M. Kojima, N. Megiddo, T. Noma, A. Yoshise (1991). *A Unified Approach to Interior Point Algorithms for Linear Complementarity Problems*. Lecture Notes in Computer Science 538. Springer, Berlin. [\[doi\]. 129](#)
- [85] S. Kräutle (2011). The semismooth Newton method for multicomponent reactive transport with minerals. *Advances in Water Resources*, 34(1), 137–151. [\[doi\]. 129](#)
- [86] J. Kruskal (1969). Two convex counterexamples: a discontinuous envelope function and a nondifferentiable nearest-point mapping. *Proceedings of the American Mathematical Society*, 23, 697–703. [20, 126](#)
- [87] H.W. Kuhn, A.W. Tucker (1951). Nonlinear programming. In J. Neyman (editor), *Proceedings of the second Berkeley Symposium on Mathematical Studies and Probability*, pages 481–492. University of California Press, Berkeley, California. [39](#)
- [88] B. Kummer (1988). Newton’s method for nondifferentiable functions. In J. Guddat, B. Bank, H. Hollatz, P. Kall, D. Klatte, B. Kummer, K. Lommatzsch, L. Tammer, M. Vlach, K. Zimmerman (editors), *Advances in Mathematical Optimization*, pages 114–125. Akademie-Verlag, Berlin. [116](#)
- [89] J. Kyparisis (1985). On uniqueness of Kuhn-Tucker multipliers in nonlinear programming. *Mathematical Programming*, 32, 242–246. [\[doi\]. 40](#)
- [90] J.B. Lasserre (2001). Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11, 796–817. [\[doi\]. 160](#)
- [91] J.B. Lasserre (2010). *Moments Positive Polynomials and Their Applications*. Imperial College Press Optimization Series 1. Imperial College Press. [160](#)
- [92] J.B. Lasserre (2015). *An Introduction to Polynomial and Semi-Algebraic Optimization*. Cambridge Texts in Applied Mathematics. Cambridge University Press. [160](#)
- [93] X. Li, D. Sun, K.-C. Toh (2018). A highly efficient semismooth Newton augmented Lagrangian method for solving Lasso problems. *SIAM Journal on Optimization*, 28(1). [\[doi\]. 126](#)
- [94] J.L. Lions, G. Stampacchia (1967). Variational inequalities. *Communication on Pure and Applied Mathematics*, 20, 493–519. [\[doi\]. 112](#)
- [95] D.G. Luenberger (1973). *Introduction to Linear and Nonlinear Programming*. Addison-Wesley, Reading, USA. [90](#)
- [96] O. Mancino, G. Stampacchia (1972). Convex programming and variational inequalities. *Journal of Optimization Theory and Applications*, 9, 3–23. [112](#)

- [97] O.L. Mangasarian, S. Fromovitz (1967). The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. *Journal of Mathematical Analysis and Applications*, 17, 37–47. [\[doi\]](#). 38
- [98] E. Marchand, T. Müller, P. Knabner (2012). Fully coupled generalised hybrid-mixed finite element approximation of two-phase two-component flow in porous media. Part II: numerical scheme and numerical results. *Computational Geosciences*, 16(3), 691–708. [\[doi\]](#). 129
- [99] E. Marchand, T. Müller, P. Knabner (2013). Fully coupled generalised hybrid-mixed finite element approximation of two-phase two-component flow in porous media. Part I: formulation and properties of the mathematical model. *Computational Geosciences*, 17(2), 431–442. [\[doi\]](#). 129
- [100] R. Mifflin (1977). Semismooth and semiconvex functions in constrained optimization. *SIAM Journal on Control and Optimization*, 15, 959–972. [\[doi\]](#). 124, 128
- [101] R.D.C. Monteiro (2003). First- and second-order methods for semidefinite programming. *Mathematical Programming*, 97, 209–244. [\[doi\]](#). 170
- [102] R.D.C. Monteiro, M. Todd (2000). Path-following methods. In H. Wolkowicz, R. Saigal, L. Vandenbergh (editors), *Handbook of Semidefinite Programming – Theory, Algorithms, and Applications*, volume 27 of *International Series in Operations Research & Management Science*, chapter 10, pages 267–306. Kluwer Academic Publishers. 170
- [103] K.G. Murty (1988). *Linear Complementarity, Linear and Nonlinear Programming* (Internet edition, prepared by Feng-Tien Yu, 1997). Heldermann Verlag, Berlin. 128
- [104] K.G. Murty, S.N. Kabadi (1987). Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39, 117–129. 90
- [105] Y.E. Nesterov, A.S. Nemirovskii (1994). *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM Studies in Applied Mathematics 13. SIAM, Philadelphia, PA, USA. 170
- [106] J.M. Ortega, W.C. Rheinboldt (1970). *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York. Reprinted in 2000 by SIAM, Classics in Applied Mathematics 30, [\[doi\]](#). 45
- [107] J.-S. Pang (1995). Complementarity problems. In R. Horst, P.M. Pardalos (editors), *Handbook of Global Optimization*, volume 2 of *Nonconvex Optimization and Its Applications*, pages 271–338. Kluwer, Dordrecht. [\[doi\]](#). 128, 129
- [108] P.A. Parrilo (2000). On a decomposition for multivariable forms via LMI methods. In *Proceedings of the American Control Conference*. 160
- [109] J.-P. Penot (1982). On regularity conditions in mathematical programming. *Mathematical Programming Study*, 19, 167–199. 80
- [110] J.-P. Penot (1994). Optimality conditions in mathematical programming and composite optimization. *Mathematical Programming*, 67, 225–245. [\[doi\]](#). 53
- [111] G. Pólya, G. Szegő (1976). *Problems and Theorems in Analysis II*. Springer-Verlag, Berlin. 170
- [112] L. Qi (1993). Convergence analysis of some algorithms for solving nonsmooth equations. *Mathematics of Operations Research*, 18, 227–244. [\[doi\]](#). 128
- [113] L. Qi, J. Sun (1993). A nonsmooth version of Newton’s method. *Mathematical Programming*, 58, 353–367. [\[doi\]](#). 128
- [114] L. Qi, X. Xiao, L. Zhang (2016). A parameter-self-adjusting Levenberg-Marquardt method for solving nonsmooth equations. *Journal of Computational Mathematics*, 34(3), 317–338. [\[doi\]](#). 127
- [115] H. Rademacher (1919). Über partielle und totale differenzierbarkeit. *I. Math. Ann.*, 89, 340–359. 31
- [116] J. Renegar (2001). *A Mathematical View of Interior-Point Methods in Convex Optimization*. MPS-SIAM Series on Optimization 3. SIAM. 155

- [117] S.M. Robinson (1974). Perturbed Kuhn-Tucker points and rates of convergence for a class of nonlinear-programming algorithms. *Mathematical Programming*, 7, 1–16. [152](#)
- [118] S.M. Robinson (1975). Stability theory for systems of inequalities, part I: linear systems. *SIAM Journal on Numerical Analysis*, 12, 754–769. [\[doi\]](#). [155](#)
- [119] S.M. Robinson (1976). Stability theory for systems of inequalities, part II: differentiable nonlinear systems. *SIAM Journal on Numerical Analysis*, 13, 497–513. [\[doi\]](#). [57](#)
- [120] S.M. Robinson (1980). Strongly regular generalized equations. *Mathematics of Operations Research*, 5, 43–62. [\[doi\]](#). [98](#), [112](#)
- [121] S.M. Robinson (1982). Generalized equations and their solutions, Part II: Applications to nonlinear programming. *Mathematical Programming Study*, 19, 200–221. [92](#), [96](#)
- [122] S.M. Robinson (1987). Local structure of feasible sets in nonlinear programming, part III: stability and sensitivity. *Mathematical Programming Study*, 30, 45–66. [\[doi\]](#). [118](#)
- [123] S.M. Robinson (1992). Normal maps induced by linear transformations. *Mathematics of Operations Research*, 17, 691–714. [113](#)
- [124] S.M. Robinson (2003). Variational conditions with smooth constraints: structure and analysis. *Mathematical Programming*, 97(1-2), 245–265. [\[doi\]](#). [99](#)
- [125] R.T. Rockafellar (1970). *Convex Analysis*. Princeton Mathematics Ser. 28. Princeton University Press, Princeton, New Jersey. [12](#), [26](#), [155](#)
- [126] R.T. Rockafellar, R. Wets (1998). *Variational Analysis*. Grundlehren der mathematischen Wissenschaften 317. Springer. [iii](#), [99](#)
- [127] A. Shapiro (1990). On concepts of directional differentiability. *Journal of Optimization Theory and Applications*, 66, 477–487. [\[doi\]](#). [143](#)
- [128] A. Shapiro (1994). Directionally nondifferentiable metric projection. *Journal of Optimization Theory and Applications*, 1, 203–204. [20](#), [126](#)
- [129] A. Shapiro (1997). On uniqueness of Lagrange multipliers in optimization problems subject to cone constraints. *SIAM Journal on Optimization*, 7, 508–518. [27](#), [74](#), [82](#)
- [130] M. Slater (1950). Lagrange multipliers revisited: a contribution to non-linear programming. Cowles Commission Discussion Paper, Math. 403. [38](#)
- [131] G. Stampacchia (1969). Variational inequalities. In *Theory and Applications of Monotone Operators (Proc. NATO Advanced Study Inst., Venice, 1968)*, pages 101–192. Edizioni Oderisi, Gubbio. [112](#)
- [132] M.J. Todd (2001). Semidefinite optimization. In *Acta Numerica 2001*, pages 515–560. Cambridge University Press. [\[doi\]](#). [170](#)
- [133] P. Tseng (1996). Global behaviour of a class of merit functions for the nonlinear complementarity problem. *Journal of Optimization Theory and Applications*, 89, 17–37. [131](#)
- [134] K. Ueda, N. Yamashita (2012). Global complexity bound analysis of the Levenberg-Marquardt method for nonsmooth equations and its application to the nonlinear complementarity problem. *Journal of Optimization Theory and Applications*, 152, 450–467. [\[doi\]](#). [127](#)
- [135] L. Vandenberghe, S. Boyd (1996). Semidefinite programming. *SIAM Review*, 38, 49–95. [\[doi\]](#). [170](#)
- [136] H. Wolkowicz, R. Saigal, L. Vandenberghe (editors) (2000). *Handbook of Semidefinite Programming – Theory, Algorithms, and Applications*, volume 27 of *International Series in Operations Research & Management Science*. Kluwer Academic Publishers. [170](#)
- [137] E.H. Zarantonello (1971). Projections on convex sets in Hilbert space and spectral theory. I: Projections on convex sets. In *Contributions to Nonlinear Functional Analysis*, pages 237–341. Academic Press, New York. Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1971. [20](#)
- [138] G. Zoutendijk (1970). Nonlinear programming, computational methods. In J. Abadie (editor), *Integer and Nonlinear Programming*, pages 37–86. North-Holland, Amsterdam. [48](#)

Index

- affine hull, 11
 - of a convex polyhedron, 26
- algorithm
 - gradient, 48
 - JN (Joseph-Newton), 102, 108
 - semismooth Newton, 126
 - SQP, 137, 150
 - SQP for (P_G) , 154
 - steepest descent, 48
- B-differential
 - definition, 118
- Banach
 - perturbation lemma, 120
- bidual cone, 15
 - expression, 17
- binary relation, 30
- C-differential
 - definition, 118
 - regular, 120
- C-function, 131
 - definition, 114
 - Fischer, 115
 - min, 115
- C-regular point, 120
- circular cone, 16
- closed function, 28
- closure, 10
- complementarity condition, 39
 - strict, *see* strict complementarity
- complementarity problem, 128
 - linear, 128
 - mixed, 129
- cone, 12
 - asymptotic, 20, 23, 26
 - bidual, *see* bidual cone
 - circular, *see* circular cone
 - critical, *see* critical cone
 - dual, *see* dual cone
 - feasible directions, 17
 - linearizing, 38, 53
 - normal, *see* normal cone
 - tangent, *see* tangent cone
- conification, 155
- constraint
 - active, 37
- constraint qualification
 - affinity (CQ-A), 38, 90
 - for representing X_E , 35
 - for representing X_{EI} , 38
 - for representing X_G , 54
 - linear independence (CQ-LI), 38, 90
 - Mangasarian-Fromovitz (CQ-MF), 38, 57, 80, 87, 88
 - Robinson (CQ-R), 57
 - Slater (CQ-S), 38
- convergence to zero, equivalently, 45
- convex optimization problem
 - (P_{comp}) , 78
 - (P_E) , 35
 - (P_{EI}) , 37
 - (P_G) , 52
- convex polyhedron, 13
 - addition, 13
 - affine hull, 26
 - dual form, 13
 - linear transformation:, 13
 - normal cone, 19
 - relative interior, 26
 - tangent cone, 19
 - upper semi-continuity of the set of active inequalities, 13
- convex set, 10
- critical cone
 - for (P_{EI}) , 84
- degree of a monomial, 160
- derivative, 28
- differentiability

- directional, *see* directional differentiability
- Fréchet, 28
- differential quotient, 24
- direction
 - critical, 84
 - feasible, 18
- directional differentiability, 22
 - composition, 143
 - in the sense of Hadamard, 143
- distance to a set, 10
- dual cone, 15
 - inclusion, 17
 - negative, 15
 - of a Cartesian product, 17
 - of a convex hull, 17
 - of a sum, 17
 - of a union, 17
 - of an intersection of cones, 17
 - self-dual, 15
- duality gap, 42

- epigraph of a function, 21
- error bound
 - and semi-stability, 103
 - definition, 58
 - Robinson for X_G , 58

- feasible direction, *see* direction
- feasible point, 32
- feasible set, 32
 - X_E of (P_E) , 35
 - X_{EI} of (P_{EI}) , 37
 - X_G of (P_G) , 52
- Finsler lemma, 158
- Fréchet, 28
- function
 - closed, 21
 - convex, 21
 - domain, 21
 - epigraph, 21
 - Lipschitz (continuous), 143
 - piecewise affine, 124
 - piecewise semismooth, 124
 - proper, 21
 - r -steep, 66
 - semismooth, 122
 - strongly semismooth, 122
 - subcontinuous, 66
 - subdifferentiable, 25
- functional inclusion, 98
 - hemi-stable solution, 111
 - semi-stable solution, 103
 - strongly regular solution, 108
- global convergence
 - of the SQP algorithm with line-search, 150
- gradient, 10

- homogenization, 155
- hull
 - affine, *see* affine hull
 - closed convex, 12
 - conical, 12
 - convex, 12
 - vector, *see* vector hull

- index set
 - active constraint, 37
 - inactive constraint, 37
- indicator function, 21
- interior, 10
- interval, 9
- isolated solution
 - of a functional inclusion, 103
 - of the optimization problem (P_G) , 92

- Kronecker symbol, 156

- Lagrangian
 - equality and inequality constrained optimization problem, 39
 - equality constrained optimization problem, 35
 - general optimization problem, 56
 - semidefinite optimization problem, 159
- Lagrangian dual
 - linear optimization, 43
- lemma
 - Banach perturbation, 120
- linear map
 - adjoint, 10
 - of a relative interior, 14
- linear preimage
 - of a relative interior, 15
- linearization
 - JN (Josephy-Newton), 102
- lower semi-continuous function, 27
- Lyusternik, 67

- machine-epsilon, 149
- matrix
 - copositive, 90

- minimizing sequence, 32
- Minkowski sum, 9
- monomial, 160
 - degree, 160
- multifunction
 - closed, 31
 - convex, 30
 - definition, 29
 - domain, 30
 - graph, 30
 - image, 30
 - inverse, 30
 - metric regular, 61
 - multiplier, 95
 - open, 61
 - range, 30
 - stationary point, 96
 - upper semi-continuous at a point, 31
- multiplier set
 - boundedness for (P_{EI}) , 40
 - boundedness for (P_G) , 74
 - boundedness for $(P_{Q,G})$, 77
 - uniqueness for (P_{EI}) , 40
- $\mathcal{N}(x)$, family of neighborhoods of x , 28
- Necessary optimality condition
 - first order
 - for (P_E) (Lagrange), 35
 - for (P_{EI}) (Karush-Kuhn-Tucker), 39
 - for (P_X) (Peano-Kantorovich), 34
 - for unconstrained optimization (Fermat), 33
- neighborhood, 10
- norm
 - dual, 144
- normal
 - map, 113
- normal cone
 - intersection, 18
 - map, 99
 - product, 18
 - to a convex set, 18
 - to a nonconvex set, 33
 - to \mathcal{S}_+^n , 158
- normal-cone-intersection, 18
- normal-cone-product, 18
- optimal multiplier
 - for (P_E) , uniqueness, 36
 - for (P_{EI}) , bounded, 40
 - for (P_{EI}) , uniqueness, 40
- optimality conditions for (P_{comp})
 - necessary of the first order (SC1), 79
 - sufficient of the first order (SC1), 79
- optimality conditions for (P_{EI})
 - necessary
 - of the second order (NC2), 88, 90
 - of the second order (NC2), semi-strong, 85, 90
 - of the second order (NC2), strong, 85, 90
 - of the second order (NC2), weak, 86, 90
 - sufficient
 - of the second order (SC2), 91
 - of the second order (SC2), semi-strong, 92
 - of the second order (SC2), strong, 92
 - of the second order (SC2), weak, 92
- optimality conditions for (P_G)
 - necessary of the first order (NC1), 55
 - sufficient of the first order (SC1), 56
- optimality conditions for $(P_{Q,G})$
 - necessary of the first order (SC1), 77
 - sufficient of the first order (SC1), 77
- optimization problem
 - bounded, 32
 - conic, 155
 - convex, *see* convex optimization problem
 - equality and inequality constrained, 37
 - equality constrained, 34
 - linear, 43
 - nonlinear, 37
 - (P_E) , 35
 - (P_{EI}) , 37, 52, 133
 - $(P_{E[l,u]})$, 52
 - (P_G) , 51
 - (P_L) , 43
 - $(P_{Q,G})$, 76
 - semidefinite, 52
 - unbounded, 32
- pairing function, 41
 - in linear optimization, 44
- polynomial
 - SOS, 161
- power set, 29
- problem
 - complementarity (P_{CP}) , 100
 - reformulation as nonsmooth equation, 114
 - functional inclusion (P_{Fi}) , 98
 - optimization, *see* optimization problem

- variational inequality
- – reformulation as nonsmooth equation, 115
- variational inequality (P_{VI}) , 113
- variational inequality (P_{VI}) , 99
- variational (P_V) , 99
- projection
 - contraction, 20
 - monotonicity, 20
- projector, 19
 - Cartesian, 30
- quadratic convergence
 - definition, 45
 - of Newton’s method for equations, 46
 - of the JN algorithm, 109–111
 - of the semismooth Newton algorithm, 126
- quadratic growth property, 91, 92
- reformulation
 - normal map, 113
- relative interior, 13
 - of a convex polyhedron, 26
- saddle-point, 42
- Schur complement, 160
- segment, 10
- semi-continuity
 - upper, *see* upper semi-continuity
- set-valued function/map/mapping, *see* multifunction
- simplex
 - unit – of \mathbb{R}^n , 11
- small o , 29
- \mathcal{S}_+^n , 11
- \mathcal{S}_{++}^n , 11
- solution
 - hemi-stable of (P_{EI}) , 138
 - improper, 43
 - semi-stable of (P_{EI}) , 138
- speed of convergence
 - linear, 45
 - quadratic, *see* quadratic convergence
 - superlinear, *see* superlinear convergence
- stability
 - of a set with respect to small perturbations, 59
- stationary point
 - for (P_E) , 35
 - for (P_{EI}) , 39
 - for (P_G) , 56
- strict complementarity, 40
 - for an \mathcal{S}_+^n -valued constraint, 82
- strong duality
 - in linear optimization, 44
- subdifferentiability, 25
- subdifferential, 25
- subgradient, 25
- superlinear convergence
 - and equivalent sequences, 45
 - definition, 45
 - Dennis and Moré criterion, 47
 - of Newton’s method for equations, 46
 - of the JN algorithm, 109–111
 - of the semismooth Newton algorithm, 126
- tangent cone
 - to a convex set, 18
 - to a nonconvex set, 33
 - to \mathcal{S}_+^n , 158
- theorem
 - mean value, 29
 - open mapping, 80
- trace of a matrix, 156
- upper semi-continuity
 - of a multifunction, 31
 - of the C-differential, 119
 - of the set of active inequalities, 13
 - of the subdifferential of a convex function, 31
- vector hull, 11
- weak duality inequality, 42
 - in linear optimization, 44
- zero of a function, 46
- Zoutendijk condition, 48