

**Grundlehren der mathematischen  
Wissenschaften**  
*A Series of Comprehensive Treatises  
in Pure and Applied Mathematics*

**Jean-Baptiste  
Claude Lemaire**

**Convex Analysis  
and Minimization  
Algorithms**

# **Grundlehren der mathematischen Wissenschaften 305**

*A Series of Comprehensive Studies in Mathematics*

## *Editors*

M. Artin S. S. Chern J. Coates J. M. Fröhlich  
H. Hironaka F. Hirzebruch L. Hörmander  
C. C. Moore J. K. Moser M. Nagata W. Schmidt  
D. S. Scott Ya. G. Sinai J. Tits M. Waldschmidt  
S. Watanabe

## *Managing Editors*

M. Berger B. Eckmann S. R. S. Varadhan

Jean-Baptiste Hiriart-Urruty  
Claude Lemaréchal

# Convex Analysis and Minimization Algorithms I

Fundamentals

With 113 Figures



Springer-Verlag Berlin Heidelberg GmbH

Jean-Baptiste Hiriart-Urruty  
Département de Mathématiques  
Université Paul Sabatier  
118, route de Narbonne  
F-31062 Toulouse, France

Claude Lemaréchal  
INRIA, Rocquencourt  
Domaine de Voluceau  
B.P. 105  
F-78153 Le Chesnay, France

## Second Corrected Printing 1996

### Library of Congress Cataloging-in-Publication Data

Hiriart-Urruty, Jean-Baptiste, 1949-  
Convex analysis and minimization algorithms / Jean-Baptiste  
Hiriart-Urruty, Claude Lemaréchal.  
p. cm. -- (Grundlehren der mathematischen Wissenschaften :  
305-306)  
"Second corrected printing"--T.p. verso.  
Includes bibliographical references (p. - ) and index.  
Contents: 1. Fundamentals -- 2. Advanced theory and bundle  
methods.

1. Convex functions. 2. Convex sets. I. Lemaréchal, Claude,  
1944-. II. Title. III. Series.  
QA331.5.H57 1993b  
515'.8--dc20

96-31946  
CIP

Mathematics Subject Classification (1991):  
26-01, 26B05, 52A41, 26A, 49K, 49M, 49-01, 93B60, 90C

ISSN 0072-7830

ISBN 978-3-642-08161-3 ISBN 978-3-662-02796-7 (eBook)  
DOI 10.1007/978-3-662-02796-7

This work is subject to copyright. All rights are reserved, whether the whole or part  
of the material is concerned, specifically the rights of translation, reprinting, reuse of  
illustrations, recitation, broadcasting, reproduction on microfilm or in any other way,  
and storage in data banks. Duplication of this publication or parts thereof is permitted  
only under the provisions of the German Copyright Law of September 9, 1965, in its  
current version, and permission for use must always be obtained from Springer-  
Verlag. Violations are liable for prosecution under the German Copyright Law.

© Springer-Verlag Berlin Heidelberg 1993

Originally published by Springer-Verlag Berlin Heidelberg New York in 1993.  
Softcover reprint of the hardcover 2nd edition 1993

Typesetting: Editing and reformatting of the authors' input files using a Springer TeX  
macro package  
SPIN: 11326120 41/3111-5 4 3 2 Printed on acid-free paper

# Table of Contents Part I

Introduction . . . . .	XV
I. Convex Functions of One Real Variable . . . . .	
1 Basic Definitions and Examples . . . . .	1
1.1 First Definitions of a Convex Function . . . . .	2
1.2 Inequalities with More Than Two Points . . . . .	6
1.3 Modern Definition of Convexity . . . . .	8
2 First Properties . . . . .	9
2.1 Stability Under Functional Operations . . . . .	9
2.2 Limits of Convex Functions . . . . .	11
2.3 Behaviour at Infinity . . . . .	14
3 Continuity Properties . . . . .	16
3.1 Continuity on the Interior of the Domain . . . . .	16
3.2 Lower Semi-Continuity: Closed Convex Functions . . . . .	17
3.3 Properties of Closed Convex Functions . . . . .	19
4 First-Order Differentiation . . . . .	20
4.1 One-Sided Differentiability of Convex Functions . . . . .	21
4.2 Basic Properties of Subderivatives . . . . .	24
4.3 Calculus Rules . . . . .	27
5 Second-Order Differentiation . . . . .	29
5.1 The Second Derivative of a Convex Function . . . . .	30
5.2 One-Sided Second Derivatives . . . . .	32
5.3 How to Recognize a Convex Function . . . . .	33
6 First Steps into the Theory of Conjugate Functions . . . . .	36
6.1 Basic Properties of the Conjugate . . . . .	38
6.2 Differentiation of the Conjugate . . . . .	40
6.3 Calculus Rules with Conjugacy . . . . .	43
II. Introduction to Optimization Algorithms . . . . .	
1 Generalities . . . . .	47
1.1 The Problem . . . . .	47
1.2 General Structure of Optimization Schemes . . . . .	50
1.3 General Structure of Optimization Algorithms . . . . .	52
2 Defining the Direction . . . . .	54

2.1 Descent and Steepest-Descent Directions . . . . .	54
2.2 First-Order Methods . . . . .	56
– One Coordinate at a Time . . . . .	56
– Euclidean Steepest Descent . . . . .	58
– General Normings . . . . .	58
2.3 Newtonian Methods . . . . .	61
2.4 Conjugate-Gradient Methods . . . . .	65
– Linear Conjugate-Gradient Method . . . . .	66
– Nonlinear Extensions . . . . .	68
3 Line-Searches . . . . .	70
3.1 General Structure of a Line-Search . . . . .	71
3.2 Designing the Test (0), (R), (L) . . . . .	74
3.3 The Wolfe Line-Search . . . . .	77
3.4 Updating the Trial Stepsize . . . . .	81
 III. Convex Sets . . . . .	87
1 Generalities . . . . .	87
1.1 Definition and First Examples . . . . .	87
1.2 Convexity-Preserving Operations on Sets . . . . .	90
1.3 Convex Combinations and Convex Hulls . . . . .	94
1.4 Closed Convex Sets and Hulls . . . . .	99
2 Convex Sets Attached to a Convex Set . . . . .	102
2.1 The Relative Interior . . . . .	102
2.2 The Asymptotic Cone . . . . .	108
2.3 Extreme Points . . . . .	110
2.4 Exposed Faces . . . . .	113
3 Projection onto Closed Convex Sets . . . . .	116
3.1 The Projection Operator . . . . .	116
3.2 Projection onto a Closed Convex Cone . . . . .	118
4 Separation and Applications . . . . .	121
4.1 Separation Between Convex Sets . . . . .	121
4.2 First Consequences of the Separation Properties . . . . .	124
– Existence of Supporting Hyperplanes . . . . .	124
– Outer Description of Closed Convex Sets . . . . .	126
– Proof of Minkowski's Theorem . . . . .	128
– Bipolar of a Convex Cone . . . . .	128
4.3 The Lemma of Minkowski-Farkas . . . . .	129
5 Conical Approximations of Convex Sets . . . . .	132
5.1 Convenient Definitions of Tangent Cones . . . . .	133
5.2 The Tangent and Normal Cones to a Convex Set . . . . .	136
5.3 Some Properties of Tangent and Normal Cones . . . . .	139

<b>IV. Convex Functions of Several Variables . . . . .</b>	<b>143</b>
1 Basic Definitions and Examples . . . . .	143
1.1 The Definitions of a Convex Function . . . . .	143
1.2 Special Convex Functions: Affinity and Closedness . . . . .	147
– Linear and Affine Functions . . . . .	147
– Closed Convex Sets . . . . .	148
– Outer Construction . . . . .	150
1.3 First Examples . . . . .	152
2 Functional Operations Preserving Convexity . . . . .	157
2.1 Operations Preserving Closedness . . . . .	158
2.2 Dilations and Perspectives of a Function . . . . .	160
2.3 Infimal Convolution . . . . .	162
2.4 Image of a Function Under a Linear Mapping . . . . .	166
2.5 Convex Hull and Closed Convex Hull of a Function . . . . .	169
3 Local and Global Behaviour of a Convex Function . . . . .	173
3.1 Continuity Properties . . . . .	173
3.2 Behaviour at Infinity . . . . .	178
4 First- and Second-Order Differentiation . . . . .	183
4.1 Differentiable Convex Functions . . . . .	183
4.2 Nondifferentiable Convex Functions . . . . .	188
4.3 Second-Order Differentiation . . . . .	190
<b>V. Sublinearity and Support Functions . . . . .</b>	<b>195</b>
1 Sublinear Functions . . . . .	197
1.1 Definitions and First Properties . . . . .	197
1.2 Some Examples . . . . .	201
1.3 The Convex Cone of All Closed Sublinear Functions . . . . .	206
2 The Support Function of a Nonempty Set . . . . .	208
2.1 Definitions, Interpretations . . . . .	208
2.2 Basic Properties . . . . .	211
2.3 Examples . . . . .	215
3 The Isomorphism Between Closed Convex Sets and Closed Sublinear Functions . . . . .	218
3.1 The Fundamental Correspondence . . . . .	218
3.2 Example: Norms and Their Duals, Polarity . . . . .	220
3.3 Calculus with Support Functions . . . . .	225
3.4 Example: Support Functions of Closed Convex Polyhedra . . . . .	234
<b>VI. Subdifferentials of Finite Convex Functions . . . . .</b>	<b>237</b>
1 The Subdifferential: Definitions and Interpretations . . . . .	238
1.1 First Definition: Directional Derivatives . . . . .	238
1.2 Second Definition: Minorization by Affine Functions . . . . .	241
1.3 Geometric Constructions and Interpretations . . . . .	243
1.4 A Constructive Approach to the Existence of a Subgradient . . . . .	247

2 Local Properties of the Subdifferential . . . . .	249
2.1 First-Order Developments . . . . .	249
2.2 Minimality Conditions . . . . .	253
2.3 Mean-Value Theorems . . . . .	256
3 First Examples . . . . .	258
4 Calculus Rules with Subdifferentials . . . . .	261
4.1 Positive Combinations of Functions . . . . .	261
4.2 Pre-Composition with an Affine Mapping . . . . .	263
4.3 Post-Composition with an Increasing Convex Function of Several Variables . . . . .	264
4.4 Supremum of Convex Functions . . . . .	266
4.5 Image of a Function Under a Linear Mapping . . . . .	272
5 Further Examples . . . . .	275
5.1 Largest Eigenvalue of a Symmetric Matrix . . . . .	275
5.2 Nested Optimization . . . . .	277
5.3 Best Approximation of a Continuous Function on a Compact Interval . . . . .	278
6 The Subdifferential as a Multifunction . . . . .	279
6.1 Monotonicity Properties of the Subdifferential . . . . .	280
6.2 Continuity Properties of the Subdifferential . . . . .	282
6.3 Subdifferentials and Limits of Gradients . . . . .	284
 VII. Constrained Convex Minimization Problems:	
Minimality Conditions, Elements of Duality Theory . . . . .	291
1 Abstract Minimality Conditions . . . . .	292
1.1 A Geometric Characterization . . . . .	293
1.2 Conceptual Exact Penalty . . . . .	298
2 Minimality Conditions Involving Constraints Explicitly . . . . .	301
2.1 Expressing the Normal and Tangent Cones in Terms of the Constraint-Functions . . . . .	303
2.2 Constraint Qualification Conditions . . . . .	307
2.3 The Strong Slater Assumption . . . . .	311
2.4 Tackling the Minimization Problem with Its Data Directly . . . . .	314
3 Properties and Interpretations of the Multipliers . . . . .	317
3.1 Multipliers as a Means to Eliminate Constraints: the Lagrange Function . . . . .	317
3.2 Multipliers and Exact Penalty . . . . .	320
3.3 Multipliers as Sensitivity Parameters with Respect to Perturbations . . . . .	323
4 Minimality Conditions and Saddle-Points . . . . .	327
4.1 Saddle-Points: Definitions and First Properties . . . . .	327
4.2 Mini-Maximization Problems . . . . .	330
4.3 An Existence Result . . . . .	333

4.4 Saddle-Points of Lagrange Functions . . . . .	336
4.5 A First Step into Duality Theory . . . . .	338
 VIII. Descent Theory for Convex Minimization:	
The Case of Complete Information . . . . .	343
1 Descent Directions and Steepest-Descent Schemes . . . . .	343
1.1 Basic Definitions . . . . .	343
1.2 Solving the Direction-Finding Problem . . . . .	347
1.3 Some Particular Cases . . . . .	351
1.4 Conclusion . . . . .	355
2 Illustration. The Finite Minimax Problem . . . . .	356
2.1 The Steepest-Descent Method for Finite Minimax Problems . . . . .	357
2.2 Non-Convergence of the Steepest-Descent Method . . . . .	363
2.3 Connection with Nonlinear Programming . . . . .	366
– Minimality Conditions . . . . .	367
– Projected Gradients in Nonlinear Programming . . . . .	367
– Projected Gradients and Steepest-Descent Directions . . . . .	369
3 The Practical Value of Descent Schemes . . . . .	371
3.1 Large Minimax Problems . . . . .	371
3.2 Infinite Minimax Problems . . . . .	373
3.3 Smooth but Stiff Functions . . . . .	374
3.4 The Steepest-Descent Trajectory . . . . .	377
– Continuous Time . . . . .	378
– Piecewise Affine Trajectories . . . . .	380
3.5 Conclusion . . . . .	383
Appendix: Notations . . . . .	385
1 Some Facts About Optimization . . . . .	385
2 The Set of Extended Real Numbers . . . . .	388
3 Linear and Bilinear Algebra . . . . .	390
4 Differentiation in a Euclidean Space . . . . .	393
5 Set-Valued Analysis . . . . .	396
6 A Bird's Eye View of Measure Theory and Integration . . . . .	399
Bibliographical Comments . . . . .	401
References . . . . .	407
Index . . . . .	415

## Table of Contents Part II

Introduction . . . . .	XV
IX. Inner Construction of the Subdifferential . . . . .	
1 The Elementary Mechanism . . . . .	1
2 Convergence Properties . . . . .	9
2.1 Convergence . . . . .	9
2.2 Speed of Convergence . . . . .	15
3 Putting the Mechanism in Perspective . . . . .	24
3.1 Bundling as a Substitute for Steepest Descent . . . . .	24
3.2 Bundling as an Emergency Device for Descent Methods . . . . .	27
3.3 Bundling as a Separation Algorithm . . . . .	29
X. Conjugacy in Convex Analysis . . . . .	
1 The Convex Conjugate of a Function . . . . .	37
1.1 Definition and First Examples . . . . .	37
1.2 Interpretations . . . . .	40
1.3 First Properties . . . . .	42
1.4 Subdifferentials of Extended-Valued Functions . . . . .	47
1.5 Convexification and Subdifferentiability . . . . .	49
2 Calculus Rules on the Conjugacy Operation . . . . .	54
2.1 Image of a Function Under a Linear Mapping . . . . .	54
2.2 Pre-Composition with an Affine Mapping . . . . .	56
2.3 Sum of Two Functions . . . . .	61
2.4 Infima and Suprema . . . . .	65
2.5 Post-Composition with an Increasing Convex Function . . . . .	69
2.6 A Glimpse of Biconjugate Calculus . . . . .	71
3 Various Examples . . . . .	72
3.1 The Cramer Transformation . . . . .	72
3.2 Some Results on the Euclidean Distance to a Closed Set . . . . .	73
3.3 The Conjugate of Convex Partially Quadratic Functions . . . . .	75
3.4 Polyhedral Functions . . . . .	76
4 Differentiability of a Conjugate Function . . . . .	79
4.1 First-Order Differentiability . . . . .	79
4.2 Towards Second-Order Differentiability . . . . .	82

XI.	Approximate Subdifferentials of Convex Functions . . . . .	91
1	The Approximate Subdifferential . . . . .	92
1.1	Definition, First Properties and Examples . . . . .	92
1.2	Characterization via the Conjugate Function . . . . .	95
1.3	Some Useful Properties . . . . .	98
2	The Approximate Directional Derivative . . . . .	102
2.1	The Support Function of the Approximate Subdifferential . . . . .	102
2.2	Properties of the Approximate Difference Quotient . . . . .	106
2.3	Behaviour of $f'_\varepsilon$ and $T_\varepsilon$ as Functions of $\varepsilon$ . . . . .	110
3	Calculus Rules on the Approximate Subdifferential . . . . .	113
3.1	Sum of Functions . . . . .	113
3.2	Pre-Composition with an Affine Mapping . . . . .	116
3.3	Image and Marginal Functions . . . . .	118
3.4	A Study of the Infimal Convolution . . . . .	119
3.5	Maximum of Functions . . . . .	123
3.6	Post-Composition with an Increasing Convex Function . . . . .	125
4	The Approximate Subdifferential as a Multifunction . . . . .	127
4.1	Continuity Properties of the Approximate Subdifferential . . . . .	127
4.2	Transportation of Approximate Subgradients . . . . .	129
XII.	Abstract Duality for Practitioners . . . . .	137
1	The Problem and the General Approach . . . . .	137
1.1	The Rules of the Game . . . . .	137
1.2	Examples . . . . .	141
2	The Necessary Theory . . . . .	147
2.1	Preliminary Results: The Dual Problem . . . . .	147
2.2	First Properties of the Dual Problem . . . . .	150
2.3	Primal-Dual Optimality Characterizations . . . . .	154
2.4	Existence of Dual Solutions . . . . .	157
3	Illustrations . . . . .	161
3.1	The Minimax Point of View . . . . .	161
3.2	Inequality Constraints . . . . .	162
3.3	Dualization of Linear Programs . . . . .	165
3.4	Dualization of Quadratic Programs . . . . .	166
3.5	Steepest-Descent Directions . . . . .	168
4	Classical Dual Algorithms . . . . .	170
4.1	Subgradient Optimization . . . . .	171
4.2	The Cutting-Plane Algorithm . . . . .	174
5	Putting the Method in Perspective . . . . .	178
5.1	The Primal Function . . . . .	178
5.2	Augmented Lagrangians . . . . .	181
5.3	The Dualization Scheme in Various Situations . . . . .	185
5.4	Fenchel's Duality . . . . .	190

<b>XIII. Methods of <math>\varepsilon</math>-Descent . . . . .</b>	<b>195</b>
1 Introduction. Identifying the Approximate Subdifferential . . . . .	195
1.1 The Problem and Its Solution . . . . .	195
1.2 The Line-Search Function . . . . .	199
1.3 The Schematic Algorithm . . . . .	203
2 A Direct Implementation: Algorithm of $\varepsilon$ -Descent . . . . .	206
2.1 Iterating the Line-Search . . . . .	206
2.2 Stopping the Line-Search . . . . .	209
2.3 The $\varepsilon$ -Descent Algorithm and Its Convergence . . . . .	212
3 Putting the Algorithm in Perspective . . . . .	216
3.1 A Pure Separation Form . . . . .	216
3.2 A Totally Static Minimization Algorithm . . . . .	219
<b>XIV. Dynamic Construction of Approximate Subdifferentials:</b>	
Dual Form of Bundle Methods . . . . .	223
1 Introduction: The Bundle of Information . . . . .	223
1.1 Motivation . . . . .	223
1.2 Constructing the Bundle of Information . . . . .	227
2 Computing the Direction . . . . .	233
2.1 The Quadratic Program . . . . .	233
2.2 Minimality Conditions . . . . .	236
2.3 Directional Derivatives Estimates . . . . .	241
2.4 The Role of the Cutting-Plane Function . . . . .	244
3 The Implementable Algorithm . . . . .	248
3.1 Derivation of the Line-Search . . . . .	248
3.2 The Implementable Line-Search and Its Convergence . . . . .	250
3.3 Derivation of the Descent Algorithm . . . . .	254
3.4 The Implementable Algorithm and Its Convergence . . . . .	257
4 Numerical Illustrations . . . . .	263
4.1 Typical Behaviour . . . . .	263
4.2 The Role of $\varepsilon$ . . . . .	266
4.3 A Variant with Infinite $\varepsilon$ : Conjugate Subgradients . . . . .	268
4.4 The Role of the Stopping Criterion . . . . .	269
4.5 The Role of Other Parameters . . . . .	271
4.6 General Conclusions . . . . .	273
<b>XV. Acceleration of the Cutting-Plane Algorithm:</b>	
Primal Forms of Bundle Methods . . . . .	275
1 Accelerating the Cutting-Plane Algorithm . . . . .	275
1.1 Instability of Cutting Planes . . . . .	276
1.2 Stabilizing Devices: Leading Principles . . . . .	279
1.3 A Digression: Step-Control Strategies . . . . .	283
2 A Variety of Stabilized Algorithms . . . . .	285
2.1 The Trust-Region Point of View . . . . .	286

2.2 The Penalization Point of View . . . . .	289
2.3 The Relaxation Point of View . . . . .	292
2.4 A Possible Dual Point of View . . . . .	295
2.5 Conclusion . . . . .	299
3 A Class of Primal Bundle Algorithms . . . . .	301
3.1 The General Method . . . . .	301
3.2 Convergence . . . . .	307
3.3 Appropriate Stepsize Values . . . . .	314
4 Bundle Methods as Regularizations . . . . .	317
4.1 Basic Properties of the Moreau-Yosida Regularization . . . . .	317
4.2 Minimizing the Moreau-Yosida Regularization . . . . .	322
4.3 Computing the Moreau-Yosida Regularization . . . . .	326
Bibliographical Comments . . . . .	331
References . . . . .	337
Index . . . . .	345

# Introduction

During the French Revolution, the writer of a project of law on public instruction complained: “Le défaut ou la disette de bons ouvrages élémentaires a été, jusqu’à présent, un des plus grands obstacles qui s’opposaient au perfectionnement de l’instruction. La raison de cette disette, c’est que jusqu’à présent les savants d’un mérite éminent ont, presque toujours, préféré la gloire d’élèver l’édifice de la science à la peine d’en éclairer l’entrée.<sup>1</sup>” Our main motivation here is precisely to “light the entrance” of the monument Convex Analysis and Minimization Algorithms. This is therefore not a reference book, to be kept on the shelf by an expert who already knows the building and can find his way through it; it is rather a book for the purpose of learning and teaching. We call above all on the intuition of the reader, and our approach is very gradual: several developments are made first in a simplified context, and then repeated in subsequent chapters at a more sophisticated level. Nevertheless, we keep constantly in mind the minimization problem suggested by A. Einstein: “Everything should be made as simple as possible, but not simpler”. Indeed, the content is by no means elementary, and will be hard for a reader not possessing a firm mastery of basic mathematical skill.

As suggested by the title, two distinct parts are involved. One, convex analysis, can be considered as an academic discipline, of a high pedagogical content, and is potentially useful to many. Minimization algorithms, on the other hand, form a much narrower subject, definitely concerning applications of mathematics, and to some extent the exclusive domain of a few specialists. Besides, we restrict ourselves to what is called nonsmooth optimization, and even more specifically to the so-called bundle algorithms. These form an important application of convex analysis, and here lies an incentive to write the present bi-disciplinary book. The theory is thus illustrated with a typical field of applications, and in return, the necessary mathematical background is thus accessible to a reader more interested by the algorithmic part. This has some consequences for the expository style: for the theoretical part, the pedagogy is based on geometric visualization of the mathematical concepts; as for minimization, only a vague knowledge of computers and numerical algorithms is assumed of the reader, which implies a rather pedestrian pace here and there.

---

<sup>1</sup>“The lack or scarcity of good, elementary books has been, until now, one of the greatest obstacles in the way of better instruction. The reason for this scarcity is that, until now, scholars of great merit have almost always preferred the glory of constructing the monument of science over the effort of lighting its entrance.” D. Guedj: *La Révolution des Savants*, Découvertes, Gallimard Sciences (1988) 130 – 131.

This dichotomous aspect emerges already in the first two chapters, which make a quick guided tour of their respective fields. Many a reader might be content with Chap. I, in which most concepts are exposed (extended-valued functions, subdifferentiability, conjugacy) in the simplest setting of univariate functions. As for Chap. II, it can be skipped by a reader familiar with classical minimization algorithms: its aim is to outline the general principles which, in our opinion, nonsmooth optimization must start from, and such a reader knows these principles.

Chapters III to VI are the instructional backbone of the work. Entirely devoted to convex analysis, they contain the basic theory, and geometric intuition is involved more than anywhere else. Chapter VII does the same thing for basic optimization theory.

Finally the last chapter of the present first part (Chap. VIII) lays down the necessary theory to develop algorithms minimizing convex functions. This chapter follows the general principles of Chap. II and serves as an illustration of basic convex analysis. On the other hand, its material is essential for a comprehension of the actual algorithms for convex (nonsmooth) optimization, to be studied in the second part.

Each chapter is presented as a “lesson”, in the sense of our old masters, treating of a given subject in its entirety. We could not completely avoid references to other chapters; but for many of them, the motivation is to suggest an intellectual link between apparently independent concepts, rather than a technical need for previous results. More than a tree, our approach evokes a spiral, made up of loosely interrelated elements.

Formally, many sections are written in smaller characters; these are not reserved to advanced material. Actually, these sections often help the reader, with illustrative examples, side remarks helping to understand a delicate point, or preparing some material to come in a subsequent chapter. Roughly speaking, they can be compared to footnotes, used to avoid interrupting the flow of the development; it can be helpful to skip them during a deeper reading, with pencil and paper. There are no formally stated exercises; but these sections in smaller characters, precisely, can often be considered as such exercises, useful to keep the reader awake.

The numbering restarts at 1 in each chapter, and chapter numbers are dropped in a cross-reference to an equation or theorem from within the same chapter. A reference of the type A.n refers to Appendix A, which recalls some theoretical background.

We thank all those, including the referees, who contributed the improvement of the manuscript by their remarks, criticisms or suggestions. Mistakes? there still must be some, of course: we just hope that they are no longer capital, and that readers will be able to detect and correct them painlessly.

Among those who helped us most, we would like to thank particularly Th. Dussaut, J.C. Gilbert, K.C. Kiwiel, S. Maurin, J.-J. Moreau, A.S. Nemirovskij, M.-R. Philippe, C.A. Sagastizábal, A. Seeger, S. Shiraishi, M. Valadier and, last but not least, the editorial and production staff of Springer-Verlag, who did a remarkably professional job. The manuscript was written on an Apple Mac+, using Microsoft Word, and CricketDraw for the pictures. It was converted into TeX with the help of “rtf2TeX”, a program written by R. Lupton at Princeton University. The final typeset version was

produced using the MathTime fonts by M. Spivak, distributed by the TeXplorators Corp. The role of OzTeX was decisive in this, and we gratefully acknowledge the technical help of W. Carlip and A. Trevorrow. Thanks and apologies are also due to Thérèse, Lydie, Sébastien, Aurélien, who had to endure our bad mood during seven years of wrestling with mathematics, computers and the English language.

Toulouse, April 1993

J.-B. Hiriart-Urruty, C. Lemaréchal

*Note about this revised printing.* Most corrections are minor; they concern misprints and other typographical details, or also informal developments. Besides, some bibliographical items have been updated and the index has been enriched.

Paris, January 1996

# I. Convex Functions of One Real Variable

**Prerequisites.** A good mastering of the following subjects: basic results from real analysis; definition and elementary properties of convex sets in  $\mathbb{R}^2$ ; elementary geometry in the affine space  $\mathbb{R}^2$ .

**Introduction.** Convex functions of a real variable form an important class of functions in the context of what is usually called real analysis. They are useful in optimization – as will be shown in this book – but also in several areas of applied mathematics, where their properties are often key ingredients to derive a priori bounds, sharp inequalities, etc.

Even though general convex functions will be studied in extenso further on, there are several reasons to devote a special chapter to the one-dimensional case.

- (i) Convexity is essentially a one-dimensional concept, since it reduces to convexity on the line joining two arbitrary points  $x$  and  $x'$ .
- (ii) For theoretical as well as algorithmic purposes, the one-dimensional trace of a convex function  $f$ , i.e. the function  $t \mapsto f(x + td)$  ( $t$  real), will have to be studied thoroughly anyway in later chapters.
- (iii) It is a good support for intuition; for example, the so-called subdifferential of a convex function can be introduced and studied very easily in the univariate case; we will also take this opportunity to introduce the concept of conjugacy operation, in this simplified setting.
- (iv) Some properties of convex functions are specific to one single variable; these properties, as well as many examples and counter-examples, will be included here.

The material contained in this chapter provides, on the one hand, sufficient background for those readers wishing to know basic properties of one-dimensional convex functions, in order to apply them in other areas of applied mathematics. On the other hand, this chapter serves as an introduction to the rest of the book; most of its results will be proved rather quickly, since they will be proved subsequently in the multi-dimensional setting. The chapter can be skipped by a reader already familiar with properties of convex functions from the viewpoint of standard real analysis. We believe, however, that our presentation may be helpful for a better understanding of the whole book.

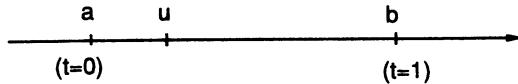
## 1 Basic Definitions and Examples

The *intervals* form the simplest instances of subsets of  $\mathbb{R}$ . We retain two among their possible definitions: a subset  $I \subset \mathbb{R}$  is an interval if and only if, whenever  $x$  and  $x'$  belong to  $I$ , one of the following properties holds:

- (i) every point between  $x$  and  $x'$  belongs to  $I$  (definition based on the natural ordering of  $\mathbb{R}$ );
- (ii) for all  $\alpha$  between 0 and 1, the point  $\alpha x + (1 - \alpha)x'$  belongs to  $I$  (definition using the vector structure of  $\mathbb{R}$ ).

The following classification of nonempty intervals is convenient:

- the compact intervals:  $I = [a, b]$  ( $a, b \in \mathbb{R}$  with  $a \leq b$ );
- the bounded but not closed intervals:  $[a, b[$ ,  $]a, b]$ ,  $]a, b[$  ( $a, b \in \mathbb{R}$ ,  $a < b$ );
- the intervals majorized but not minorized – resp. minorized but not majorized:  $] -\infty, b]$  and  $] -\infty, b[$  ( $b \in \mathbb{R}$ ) – resp.  $[a, +\infty[$  and  $]a, +\infty[$  ( $a \in \mathbb{R}$ );
- the only interval neither majorized nor minorized, namely  $\mathbb{R}$  itself.



**Fig. 1.0.1.** Parametrization of an interval

Bounded intervals will also be called segments, or line-segments. The following parametric representation, illustrated on Fig. 1.0.1, is classical for a point  $u \in ]a, b[$ :

$$u = \alpha b + (1 - \alpha)a = a + \alpha(b - a) \quad \text{with} \quad \alpha = \frac{u - a}{b - a} \in ]0, 1[. \quad (1.0.1)$$

Finally, we recall basic definitions for a function  $f : D \rightarrow \mathbb{R}$ .

**Definition 1.0.1** The *graph* of  $f$  is the subset of  $D \times \mathbb{R}$

$$\text{gr } f := \{(x, r) : x \in D \text{ and } r = f(x)\}.$$

The *epigraph* of  $f$  is “everything that lies above the graph”:

$$\text{epi } f := \{(x, r) : x \in D \text{ and } r \geq f(x)\}.$$

The strict epigraph is defined likewise, with “ $\geq$ ” replaced by “ $>$ ”. □

Thus,  $\text{epi } f$  is a juxtaposition of closed non-majorized intervals in  $\mathbb{R}$ , of the form  $[a, +\infty[$  with  $a = f(x)$ . In convex analysis, asymmetry arises naturally: the “hypograph” of a function presents no additional interest.

## 1.1 First Definitions of a Convex Function

The very first definition of a convex function is as follows:

**Definition 1.1.1 (Analytical)** Let  $I$  be a nonempty interval of  $\mathbb{R}$ . A function  $f : I \rightarrow \mathbb{R}$  is said to be *convex on  $I$*  when

$$f(\alpha x + (1 - \alpha)x') \leq f(x) + (1 - \alpha)f(x') \quad (1.1.1)$$

for all pairs of points  $(x, x')$  in  $I$  and all  $\alpha \in ]0, 1[$ .

It is said to be *strictly convex* when strict inequality holds in (1.1.1) if  $x \neq x'$ .  $\square$

The geometric meaning of convexity is clear: consider on Fig. 1.1.1 the segment  $P_x P_{x'}$  joining in  $\mathbb{R}^2$  the point  $P_x = (x, f(x))$  to the point  $P_{x'} = (x', f(x'))$ . To say that  $f$  is convex is to say that, for all  $x, x'$  in  $I$  and all  $u$  in  $]x, x'[$ , the point  $P_u = (u, f(u))$  of  $\text{gr } f$  lies below the segment  $P_x P_{x'}$  (without loss of generality, we assume  $x < x'$ ).

Once the geometrical meaning of convexity is understood, the following equivalent characterization is easily derived:

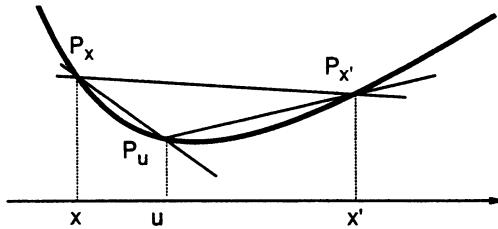


Fig. 1.1.1. The fundamental property of a convex epigraph

**Definition 1.1.2 (Geometrical)** Let  $I$  be a nonempty interval of  $\mathbb{R}$ . A function  $f : I \rightarrow \mathbb{R}$  is convex on  $I$  if and only if  $\text{epi } f$  is a convex subset of  $\mathbb{R}^2$ . (We recall the definition of a convex set in  $\mathbb{R}^2$ : it is a set  $C$  such that, if the points  $P$  and  $P'$  are in  $C$ , then the segment joining  $P$  to  $P'$  is also in  $C$ ).

Equivalently, a function is convex when its strict epigraph is convex.  $\square$

Figure 1.1.1 suggests that, since  $u$  lies between  $x$  and  $x'$  and  $P_u$  lies below  $P_x P_{x'}$ , the slope of  $P_x P_u$  (rather: of the line joining  $P_x$  and  $P_u$ ) is smaller than the slope of  $P_x P_{x'}$ , which itself is smaller than the slope of  $P_u P_{x'}$ . The next result from elementary geometry in the affine space  $\mathbb{R}^2$  clarifies the argument.

**Proposition 1.1.3** Let  $P_x = (x, y)$ ,  $P_u = (u, v)$  and  $P_{x'} = (x', y')$  be three points in  $\mathbb{R}^2$ , with  $u \in ]x, x'[$ . Then the following three properties are equivalent:

- (i)  $P_u$  is below  $P_x P_{x'}$ ;
- (ii)  $\text{slope}(P_x P_u) \leq \text{slope}(P_x P_{x'})$ ;
- (iii)  $\text{slope}(P_x P_{x'}) \leq \text{slope}(P_u P_{x'})$ .

PROOF. Property (i) means  $v \leq y + \frac{y'-y}{x'-x}(u-x)$ ; this implies  $\frac{v-y}{u-x} \leq \frac{y'-y}{x'-x}$ , which is (ii); and so on.  $\square$

Translated into the graph-language, (ii) and (iii) above mean

$$\frac{f(u) - f(x)}{u - x} \leq \frac{f(x') - f(x)}{x' - x} \leq \frac{f(x') - f(u)}{x' - u}, \quad (1.1.2)$$

which can also be obtained via the representation (1.0.1) of  $u \in ]x, x'[:$  plugging it into the definition (1.1.1) of convexity gives

$$\begin{aligned} f(u) &\leq \frac{x' - u}{x' - x} f(x) + \frac{u - x}{x' - x} f(x') = \\ &= f(x) + \frac{f(x') - f(x)}{x' - x} (u - x) = f(x') + \frac{f(x) - f(x')}{x' - x} (x' - u), \end{aligned}$$

and this displays the connection between (1.1.1) and mean-value relations such as (1.1.2).

Combining the geometric definition of a convex function with the equivalence stated in Proposition 1.1.3 gives the following *characterization* of convexity:

**Proposition 1.1.4 (Criterion of Increasing Slopes)** *Let  $I$  be a nonempty interval of  $\mathbb{R}$ . A function  $f : I \rightarrow \mathbb{R}$  is convex on  $I$  if and only if, for all  $x_0 \in I$ , the slope-function*

$$x \mapsto \frac{f(x) - f(x_0)}{x - x_0} =: s(x) \quad (1.1.3)$$

*is increasing on  $I \setminus \{x_0\}$ .*  $\square$

Knowing that every  $P_u = (u, f(u)) \in \text{gr } f$  lies below the line  $P_x P_{x'}$  when  $u \in ]x, x'[,$  what happens outside this last interval? Proposition 1.1.4 implies that, for  $v \notin [x, x'], P_v$  lies *above* the line  $P_x P_{x'}.$  To see it, exchange  $u$  and  $x'$  on Fig. 1.1.1.

If  $\varphi$  is an increasing function on a segment  $[a, b],$  the convexity of the function

$$[a, b] \ni x \mapsto f(x) := \int_a^x \varphi(u) du$$

is easily established from Definition 1.1.1. Take  $\alpha \in ]0, 1[,$  and  $a \leq x < x' \leq b;$  set  $x'' := \alpha x + (1 - \alpha)x'$  and compute

$$f(x'') - \alpha f(x) - (1 - \alpha)f(x') =: \Phi$$

(which must be nonpositive). We have

$$\Phi = \alpha \int_x^{x''} \varphi(u) du + (\alpha - 1) \int_{x''}^{x'} \varphi(u) du$$

and, using the monotonicity of  $\varphi,$  we get

$$\Phi \leq \alpha \varphi(x'') (\alpha - 1)(x - x') + \alpha(\alpha - 1) \varphi(x'') (x' - x) = 0.$$

We mention some other examples.

**Examples 1.1.5** For  $r > 0$ , draw the graph of the function whose value at  $x$  is

$$f_{1/r}(x) := \begin{cases} \frac{1}{2}rx^2 & \text{for } |x| \leq 1/r, \\ |x| - \frac{1}{2r} & \text{for } |x| \geq 1/r, \end{cases} \quad (1.1.4)$$

to notice that it is convex on  $\mathbb{R}$ . Look at what happens when  $r \rightarrow +\infty$ ; when  $r \downarrow 0$ .

- The function  $x \mapsto |x|$  is also convex on  $\mathbb{R}$ .
- The function  $x \mapsto f(x) := \sqrt{1+x^2}$  is convex on  $\mathbb{R}$ .
- If  $\varphi : [0, 1] \rightarrow \mathbb{R}$  is continuously differentiable, remember that

$$L[\varphi] := \int_0^1 f(\varphi'(u))du = \int_0^1 \sqrt{1+\varphi'^2(u)}du$$

is the length of the curve  $\{u, \varphi(u)\}_{u \in [0, 1]}$ . The convexity of  $f$  ensures the “convexity” of  $L$ :

$$L[\alpha\varphi + (1-\alpha)\psi] \leq \alpha L[\varphi] + (1-\alpha)L[\psi] \quad \text{for } \alpha \in ]0, 1[,$$

a property very useful if one wishes to minimize  $L$ . □

Up to now, the tools we have on hand to establish convexity are 1.1.1, 1.1.2 and 1.1.4. They are still rather coarse (§5 will give more in terms of differential calculus) but the criterion of increasing slopes can be useful. An example is the following important result.

**Theorem 1.1.6** *Let  $f$  be defined on  $]0, +\infty[$ . Then the function*

$$0 < x \mapsto g(x) := xf(1/x)$$

*is convex on  $]0, +\infty[$  if and only if  $f$  also is convex on  $]0, +\infty[$ .*

PROOF. Suppose  $f$  is convex on  $]0, +\infty[$ ; let  $x_0 > 0$  and consider the slope-function

$$s_g(x) := \frac{g(x) - g(x_0)}{x - x_0} = \frac{xf(1/x) - x_0f(1/x_0)}{x - x_0},$$

defined on  $]0, +\infty[\setminus\{x_0\}$ . We have

$$\begin{aligned} s_g(x) &= \frac{x - x_0}{x - x_0}f(1/x_0) + \frac{x}{x - x_0}[f(1/x) - f(1/x_0)] \\ &= f(1/x_0) - \frac{1}{x_0} \frac{f(1/x) - f(1/x_0)}{1/x - 1/x_0} = f(1/x_0) - \frac{1}{x_0}s_f(1/x). \end{aligned}$$

When  $x$  increases,  $1/x$  decreases,  $s_f(1/x)$  decreases (criterion 1.1.4 of increasing slopes) and  $s_g(x)$  therefore increases:  $g$  is convex. The “only if” part clearly follows if we observe that  $xg(1/x) = f(x)$ . □

For example, if we know that the functions  $-\log x$  and  $\exp x$  are convex on  $]0, +\infty[$ , we immediately deduce the convexity of the functions  $x \log x$  and  $x \exp 1/x$ .

## 1.2 Inequalities with More Than Two Points

An essential feature of the basic inequality (1.1.1) is that it can be generalized to more than two points.

**Theorem 1.2.1** *Let  $I$  be a nonempty interval of  $\mathbb{R}$  and  $f$  be convex on  $I$ . Then, for any collection  $\{x_1, \dots, x_k\}$  of points in  $I$  and any collection of numbers  $\{\alpha_1, \dots, \alpha_k\}$  satisfying*

$$\alpha_i \geq 0 \text{ for } i = 1, \dots, k \quad \text{and} \quad \sum_{i=1}^k \alpha_i = 1, \quad (1.2.1)$$

*Jensen's inequality holds (in summation form):*

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

PROOF. Consider first  $k = 2$ . The relation is trivial if  $\alpha_1$  or  $\alpha_2$  is zero; if not, it is just (1.1.1).

Now, suppose inductively that the relation is true for  $k - 1$ ; let a collection  $\{x_i\}$  and  $\{\alpha_i\}$  be as in (1.2.1). If  $\alpha_k$  is 0 or 1, there is nothing to prove. If not, set

$$\bar{\alpha} := \sum_{i=1}^{k-1} \alpha_i \in ]0, 1[ \quad (\alpha_k = 1 - \bar{\alpha} \in ]0, 1[),$$

$$\bar{\alpha}_i := \frac{\alpha_i}{\bar{\alpha}} \quad \text{for } i = 1, \dots, k-1 \quad \left( \bar{\alpha}_i \geq 0, \sum_{i=1}^{k-1} \bar{\alpha}_i = 1 \right),$$

so that

$$\sum_{i=1}^k \alpha_i x_i = \bar{\alpha} \sum_{i=1}^{k-1} \bar{\alpha}_i x_i + (1 - \bar{\alpha}) x_k.$$

In this last relation, the point  $\bar{x} := \sum_{i=1}^{k-1} \bar{\alpha}_i x_i$  is in  $I$  (it is between  $\min_i x_i$  and  $\max_i x_i$ ). We can therefore apply (1.1.1) to obtain

$$f\left(\sum_{i=1}^k \alpha_i x_i\right) \leq \bar{\alpha} f(\bar{x}) + (1 - \bar{\alpha}) f(x_k) = \bar{\alpha} f(\bar{x}) + \alpha_k f(x_k).$$

Then the result follows from the induction assumption applied to  $\bar{x}$ :

$$\bar{\alpha} f(\bar{x}) \leq \bar{\alpha} \sum_{i=1}^{k-1} \bar{\alpha}_i f(x_i) = \sum_{i=1}^{k-1} \alpha_i f(x_i). \quad \square$$

The set described by (1.2.1) is called the *unit simplex* of  $\mathbb{R}^k$ . A collection of  $\alpha_i$ 's satisfying (1.2.1) is called a set of *convex multipliers* and the corresponding  $x = \sum_{i=1}^k \alpha_i x_i$  is a *convex combination* of the  $x_i$ 's.

We claim that most useful inequalities between real numbers are consequences of the above Jensen inequality, even if it is not always easy to discover the underlying convex function. Let us give some typical examples.

**Example 1.2.2** Suppose we know that the function  $-\log x$  is convex on  $]0, +\infty[$ :

$$-\log \left( \sum_{i=1}^k \alpha_i x_i \right) \leq -\sum_{i=1}^k \alpha_i \log x_i = -\log \left( \prod_{i=1}^k x_i^{\alpha_i} \right)$$

for all positive  $x_i$  and  $\alpha = (\alpha_1, \dots, \alpha_k)$  in the unit simplex. Then the monotonicity of the exponential gives

$$\prod_{i=1}^k x_i^{\alpha_i} \leq \sum_{i=1}^k \alpha_i x_i.$$

Furthermore, since the function  $x \log x$  is convex, a similar calculation gives

$$\left( \sum_{i=1}^k \alpha_i x_i \right)^{\sum_{i=1}^k \alpha_i x_i} \leq \prod_{i=1}^k x_i^{\alpha_i x_i}. \quad \square$$

**Example 1.2.3** Sometimes, the use of Jensen's inequality may call for an  $f$  which is hard to find. Below, we present without details some relations, more or less classical, giving for each of them the convex "generator" (whose convexity will follow from the more refined criteria in §5).

– Let  $\alpha$  be in the unit simplex and  $\{x_i\}$  in  $]0, 1]$ . Then

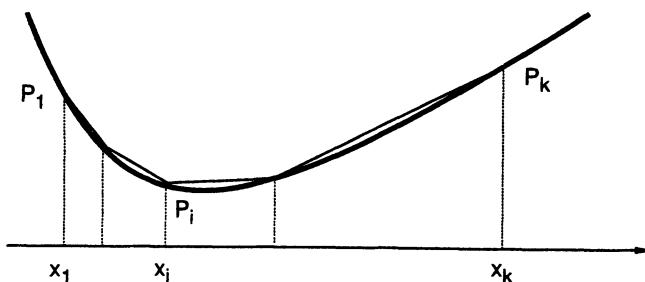
$$\sum_{i=1}^k \frac{\alpha_i}{1+x_i} \leq \left( 1 + \prod_{i=1}^k x_i^{\alpha_i} \right)^{-1}$$

(use the function  $y \mapsto -1/(1+e^{-y})$  on  $[0, +\infty[$  and consider  $y_i = -\log x_i$ ).

– Let  $\alpha$  be in the unit simplex,  $\{x_i\}$  and  $\{y_i\}$  be positive. Then

$$\prod_{i=1}^k x_i^{\alpha_i} + \prod_{i=1}^k y_i^{\alpha_i} \leq \sum_{i=1}^k (x_i + y_i)^{\alpha_i}$$

(use the function  $u \mapsto \log(1+\exp u)$  on  $\mathbb{R}$  and consider  $u_i = \log y_i - \log x_i$ ).  $\square$



**Fig. 1.2.1.** Inner approximation of a convex epigraph

The criterion of increasing slopes, illustrated in Fig. 1.1.1, also lends itself to generalization. We simply refer to Fig. 1.2.1: if  $x_1 < x_2 < \dots < x_k$  lie in the interval  $I$  where  $f$  is convex, the slopes  $[f(x_{i+1}) - f(x_i)]/(x_{i+1} - x_i)$  increase with  $i$ .

Denote by  $P_i$  the point  $(x_i, f(x_i))$  on  $\text{gr } f$  and consider the piecewise affine function  $\check{f}$ , whose graph is the sequence of segments  $[P_i, P_{i+1}]$ . This graph is above  $\text{gr } f$ :

$$\check{f}(x) \geq f(x) \quad \text{for all } x \in I.$$

It follows for example that, when approximating the integral of  $f$  by the trapezoidal rule, the resulting error has a definite sign.

### 1.3 Modern Definition of Convexity

When dealing with convexity, it is convenient to consider a function  $f$  as being defined *on the whole space*  $\mathbb{R}$ , by allowing the value  $+\infty$  for  $f(x)$ . Until now, convexity involved a pair  $(I, f)$ , where  $I$  was a nonempty interval and  $f$  a function from  $I$  to  $\mathbb{R}$ , satisfying (1.1.1) on  $I$ . We can extend such an  $f$  beyond  $I$  via the function

$$f_e(x) := \begin{cases} f(x) & \text{for } x \in I, \\ +\infty & \text{for } x \notin I. \end{cases}$$

This *extended-valued* function  $f_e$  sends  $\mathbb{R}$  to the set  $\mathbb{R} \cup \{+\infty\}$  (extended calculus is introduced in the appendix §A.2); of course, the value  $+\infty$  has been carefully selected: it is the only way to preserve the relation of definition (1.1.1) outside  $I$ . From now on and without explicit mention, all (potentially) convex functions will be extended-valued: the subscript “e” will therefore be dropped and the definitions of §1.1 are accordingly replaced as follows:

**Definition 1.3.1** A function  $f : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ , not identically equal to  $+\infty$ , is said to be convex when the inequality in  $\mathbb{R} \cup \{+\infty\}$

$$f(\alpha x + (1 - \alpha)x') \leq \alpha f(x) + (1 - \alpha)f(x') \tag{1.3.1}$$

holds for all pairs of points  $(x, x')$  in  $\mathbb{R}$  and all  $\alpha \in ]0, 1[$ .

Equivalently, it is a function whose epigraph is a nonempty convex set in  $\mathbb{R} \times \mathbb{R}$ .

The set of such functions is denoted by  $\text{Conv } \mathbb{R}$ .  $\square$

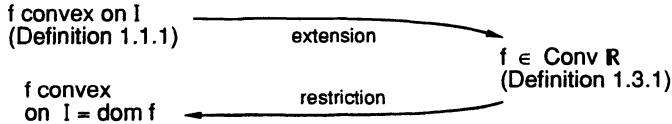
It is on purpose that the somewhat pathological function  $f \equiv +\infty$  is eliminated from  $\text{Conv } \mathbb{R}$ ; it presents no interest (note: its graph and epigraph are empty). The new definition alleviates notation, in that the interval  $I$  can be dropped when not needed. However, it has not suddenly become totally useless, and the concept must not be forgotten:

**Definition 1.3.2** The *domain* of  $f \in \text{Conv } \mathbb{R}$  is the nonempty set

$$\text{dom } f := \{x \in \mathbb{R} : f(x) \in \mathbb{R}\}. \quad \square$$

Naturally,  $\text{dom } f$  is an interval, say  $I$ , and  $f$  is after all nothing more than a convex function on  $I$  (in the sense of Definition 1.1.1). In short, two simple operations are involved, as displayed in Fig. 1.3.1.

The usefulness of Definition 1.3.1 is more than notational; it is especially convenient when optimization is involved. Let us give three examples to illustrate this.



**Fig. 1.3.1.** “Classical” and extended-valued convex functions

- Let  $x$  be a real parameter and consider the simple optimization problem

$$\inf \{-y : y^2 \leq x\}. \quad (1.3.2)$$

It is meaningless if  $x < 0$  but, for  $x \geq 0$ , the optimal value is  $-\sqrt{x}$ , a convex function of  $x$ . In view of the convention  $\inf \emptyset = +\infty$ , we do have a convex function in the sense of Definition 1.3.1. It is good to know that problems of the type (1.3.2) yield convex functions of  $x$  (this will be confirmed in Chap. IV), even though they may not be meaningful for all values of  $x$ .

- Associated to a given  $f$  is the so-called *conjugate function*

$$\mathbb{R} \ni x \mapsto \sup \{xy - f(y) : y \in \mathbb{R}\}.$$

Here again, the values of  $x$  for which the supremum is finite are not necessarily known beforehand. This supremum is thus an extended-valued function of  $x$ , a function which turns out to be of utmost importance.

- Suppose that a function  $g$ , convex on  $I$ , must be minimized on some nonempty subinterval  $C \subset I$ . The constraint  $x \in C$  can be included in the objective function by setting

$$f(x) := \begin{cases} g(x) & \text{if } x \in C, \\ +\infty & \text{if not.} \end{cases}$$

The resulting  $f$  is in  $\text{Conv } \mathbb{R}$  and minimizing it (on the whole of  $\mathbb{R}$ ) is just equivalent to the original problem.

**Remark 1.3.3** The price to pay when accepting  $f(x) = +\infty$  is alluded to in §A.2: some care must be exercised when doing algebraic manipulations; essentially, multiplications of function-values by nonpositive numbers should be avoided whenever possible. This was done already in (1.1.1) or (1.3.1), where the requirement  $\alpha \in ]0, 1[$  (rather than  $\alpha \in [0, 1]$ ) was not totally innocent.  $\square$

## 2 First Properties

### 2.1 Stability Under Functional Operations

In this section, we list some of the operations which can be proved to preserve convexity, simply in view of the definitions themselves.

**Proposition 2.1.1** *Let  $f_1, \dots, f_m$  be  $m$  convex functions and  $t_1, \dots, t_m$  be positive numbers. If there exists  $x_0$  such that  $f_j(x_0) < +\infty$ ,  $j = 1, \dots, m$ , then the function  $f := \sum_{j=1}^m t_j f_j$  is in  $\text{Conv } \mathbb{R}$ .*

PROOF. Immediate from the relation of definition (1.3.1).  $\square$

Note the precaution above: when adding two functions  $f_1$  and  $f_2$ , we have to make sure that their sum is not identically  $+\infty$ , i.e. that  $\text{dom } f_1 \cap \text{dom } f_2 \neq \emptyset$ .

**Proposition 2.1.2** *Let  $\{f_j\}_{j \in J}$  be a family of convex functions. If there exists  $x_0 \in \mathbb{R}$  such that  $\sup_{j \in J} f_j(x_0) < +\infty$ , then the function  $f := \sup_{j \in J} f_j$  is in  $\text{Conv } \mathbb{R}$ .*

PROOF. Observe that the epigraph of  $f$  is the intersection over  $J$  of the convex sets  $\text{epi } f_j$ .  $\square$

The minimum of two convex functions is certainly not convex in general (draw a picture). However, the inf-operation does preserve convexity in a slightly more elaborate setting.

**Proposition 2.1.3** *Let  $f_1$  and  $f_2$  be convex and set for all  $x \in \mathbb{R}$*

$$\begin{aligned} f(x) := (f_1 \downarrow f_2)(x) &:= \inf\{f_1(x_1) + f_2(x_2) : x_1 + x_2 = x\} \\ &= \inf\{f_1(y) + f_2(x - y) : y \in \mathbb{R}\}. \end{aligned} \quad (2.1.1)$$

If there exist two real numbers  $s_0$  and  $r_0$  such that

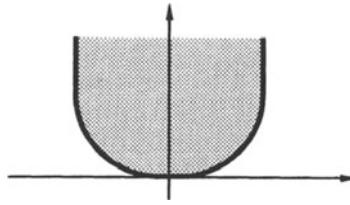
$$f_j(x) \geq s_0 x - r_0 \quad \text{for } j = 1, 2 \text{ and all } x \in \mathbb{R}$$

(in other words, the affine function  $x \mapsto s_0 x - r_0$  minorizes  $f_1$  and  $f_2$ ), then  $f \in \text{Conv } \mathbb{R}$ .

**EXPLANATION.** The domain of  $f$  in (2.1.1) is  $\text{dom } f_1 + \text{dom } f_2$ : by construction,  $f(x) < +\infty$  if  $x_1$  and  $x_2$  can be found such that  $x_1 + x_2 = x$  and  $f_1(x_1) + f_2(x_2) < +\infty$ . On the other hand,  $f(x)$  is minorized by  $s_0 x - 2r_0 > -\infty$  for all  $x$ . Now, an algebraic proof of convexity, based on (1.3.1), would be cumbersome. The key is actually to realize that the strict epigraph of  $f$  is the sum (in  $\mathbb{R}^2$ ) of the strict epigraphs of  $f_1$  and  $f_2$ : see Definitions 1.0.1 and 1.1.2.  $\square$

The operation described by (2.1.1) is called the *infimal convolution* of  $f_1$  and  $f_2$ . It is admittedly complex but important and will be encountered on many occasions. Let us observe right here that it corresponds to the (admittedly simple) addition of epigraphs – barring some technicalities. It is a good exercise to visualize the infimal convolution of an arbitrary convex  $f_1$  and

- $f_2(x) = r$  if  $x = 0$ ,  $+\infty$  if not (shift  $\text{epi } f_1$  vertically by  $r$ );
- $f_2(x) = 0$  if  $x = x_0$ ,  $+\infty$  if not (horizontal shift);
- $f_2(x) = 0$  if  $|x| \leq r$ ,  $+\infty$  if not (horizontal smear);
- $f_2(x) = sx - r$  (it is  $\text{gr } f_2$  that wins);



**Fig. 2.1.1.** The ball-pen function

- $f_2(x) = 1 - \sqrt{1 - x^2}$  for  $x \in [-1, +1]$  (the “ball-pen function” of Fig. 2.1.1); translate the bottom of the ball-pen (the origin of  $\mathbb{R}^2$ ) to each point in  $\text{gr } f_1$ ;
- $f_2(x) = 1/2 x^2$  (similar operation).

**Remark 2.1.4** The classical (integral) convolution between two functions  $F_1$  and  $F_2$  is

$$(F_1 * F_2)(x) := \int_{\mathbb{R}} F_1(y) F_2(x - y) dy \quad \text{for all } x \in \mathbb{R}.$$

For nonnegative functions, we can consider the “convolution of order  $p$ ” ( $p > 0$ ):

$$(F_1 *_p F_2)(x) := \left\{ \int_{\mathbb{R}} [F_1(y) F_2(x - y)]^p dy \right\}^{1/p} \quad \text{for all } x \in \mathbb{R}.$$

It is reasonable to claim that this integral converges to  $\sup_y F_1(y) F_2(x - y)$  when  $p \rightarrow +\infty$ . Now take  $F_i := \exp(-f_i)$ ,  $i = 1, 2$ ; we have

$$(F_1 *_{\infty} F_2)(x) = \sup_y e^{-f_1(y) - f_2(x - y)} = e^{-\inf_y [f_1(y) + f_2(x - y)]}.$$

Thus, the infimal convolution appears as a “convolution of infinite order”, combined with an exponentiation.

## 2.2 Limits of Convex Functions

**Proposition 2.2.1** Consider a sequence  $\{f_k\}_{k \in \mathbb{N}}$  of functions in  $\text{Conv } \mathbb{R}$ . Assume that, when  $k \rightarrow +\infty$ ,  $\{f_k\}$  converges pointwise (in  $\mathbb{R} \cup \{+\infty\}$ ) to a function  $f : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$  which is not identically  $+\infty$ . Then  $f \in \text{Conv } \mathbb{R}$ .

PROOF. Apply (1.3.1) to  $f_k$  and let  $k \rightarrow +\infty$ . □

**Remark 2.2.2** The interval  $\text{dom } f_k$  may depend on  $k$ . Special attention is often paid to the behaviour of  $f_k$  on some fixed interval  $I$  contained in  $\text{dom } f_k$  for all  $k$ . If, in addition,  $I$  is contained in the domain of the limit-function  $f$ , then a stronger result can be proved: the convergence of  $f_k$  to  $f$  is *uniform* on any compact subinterval of  $\text{int } I$ . □

It is usual in analysis to approximate a given function  $f$  by a sequence of more “regular” functions  $f_k$ . In the presence of convexity, we give two examples of regularization, based on the convolution operation.

Our first example is classical. Choose a “kernel function”  $K : \mathbb{R} \rightarrow \mathbb{R}^+$ , which is continuous, vanishes outside some compact interval, and is such that  $\int_{\mathbb{R}} K(y)dy = 1$ ; define for positive integer  $k$  the function

$$\mathbb{R} \ni y \mapsto K_k(y) := kK(ky).$$

Given a function  $f$  (to simplify notation, suppose  $\text{dom } f = \mathbb{R}$ ), the convolution

$$f_k(x) := f * K_k = \int_{\mathbb{R}} f(x-y)K_k(y)dy \quad (2.2.1)$$

is an approximation of  $f$ , and its smoothness properties just depend on those of  $K$ . If  $K \in C^\infty(\mathbb{R})$ , a  $C^\infty$  regularization is obtained; such is the case for example with

$$K(y) := c \exp \frac{1}{y^2 - 1} \quad \text{for } |y| < 1 \quad (0 \text{ outside}),$$

where  $c > 0$  is chosen so that  $K$  has integral 1.

**Proposition 2.2.3** *Let  $\{K_k\}_{k \in \mathbb{N}}$  be a sequence of  $C^\infty$  kernel-functions as defined above, and let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be convex. Then  $f_k$  of (2.2.1) is a  $C^\infty$  convex function on  $\mathbb{R}$ , and  $\{f_k\}$  converges to  $f$  uniformly on any compact subset of  $\mathbb{R}$ .*

PROOF. The convexity of  $f_k$  comes immediately from the analytical definition (1.1.1). The  $C^\infty$ -property of  $f_k$  and the convergence result of  $\{f_k\}$  to  $f$  are classical in real analysis.  $\square$

Another type of regularization uses the infimal convolution with a *convex* kernel  $K_k$ . It plays an important role in convex analysis and optimization, for both theoretical and algorithmic aspects. We give two examples of kernel functions:

$$K_k(y) := \frac{1}{2}ky^2 \quad \text{and} \quad K_k(y) := k|y|,$$

which have the following effects (the proofs are omitted and will be given later in §XV.4.1 and §XI.3.4):

**Proposition 2.2.4** *Let  $f \in \text{Conv } \mathbb{R}$ . For all positive  $k$ , define*

$$f_{(k)}(x) := \inf \left\{ f(y) + \frac{1}{2}k(x-y)^2 : y \in \mathbb{R} \right\}; \quad (2.2.2)$$

*then:*

- (i)  $f_{(k)}$  is convex from  $\mathbb{R}$  to  $\mathbb{R}$  and  $f_{(k)}(x) \leq f_{(k+1)}(x) \leq f(x)$  for all  $x \in \mathbb{R}$ ;
- (ii) if  $x_0$  minimizes  $f$  on  $\mathbb{R}$ , it also minimizes  $f_{(k)}$  and then  $f_{(k)}(x_0) = f(x_0)$ ; the converse is true whenever  $x_0$  is in the interior of  $\text{dom } f$ ;
- (iii)  $f_{(k)}$  is differentiable and its derivative is Lipschitz-continuous:

$$|f'_{(k)}(x_1) - f'_{(k)}(x_2)| \leq k|x_1 - x_2| \quad \text{for all } (x_1, x_2) \in \mathbb{R} \times \mathbb{R};$$

- (iv) except possibly on the boundary of  $\text{dom } f$ ,  $\{f_{(k)}\}$  converges pointwise to  $f$  when  $k \rightarrow +\infty$ .

For  $k$  large enough, define

$$f_{[k]}(x) := \inf \{f(y) + k|x - y| : y \in \mathbb{R}\}; \quad (2.2.3)$$

then:

- (j)  $f_{[k]}$  is convex from  $\mathbb{R}$  to  $\mathbb{R}$  and  $f_{[k]}(x) \leq f_{[k+1]}(x) \leq f(x)$  for all  $x \in \mathbb{R}$ ;
- (jj) if  $x_0 \in \text{int dom } f$ , then  $f_{[k]}(x_0) = f(x_0)$  for  $k$  large enough;
- (jjj)  $f_{[k]}$  is Lipschitz-continuous:

$$|f_{[k]}(x_1) - f_{[k]}(x_2)| \leq k|x_1 - x_2| \quad \text{for all } (x_1, x_2) \in \mathbb{R} \times \mathbb{R}. \quad \square$$

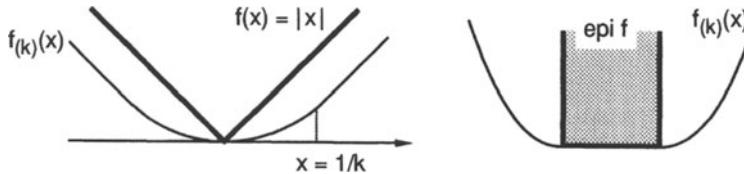


Fig. 2.2.1. Moreau-Yosida  $C^1$  regularizations

Replacing  $f$  by  $f_{(k)}$  of (2.2.2) is called *Moreau-Yosida regularization*. It yields  $C^1$ -smoothness, without essentially changing the set of minimizers; note also that the function to be minimized in (2.2.2) is strictly convex (and even better: so-called strongly convex). It is not too difficult to work out the calculations when  $f(x) = |x|$ : the result is the function of (1.1.4), illustrated on Fig. 2.2.1. It has a continuous derivative, as claimed in (iii), but no second derivative at  $x = \pm 1/k$ . The right part of the picture shows the effect of the same regularization on another function.

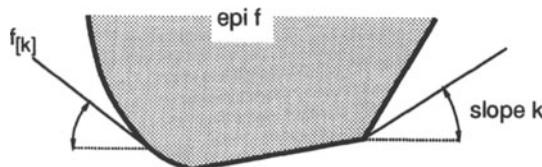


Fig. 2.2.2. A Lipschitzian regularization

Note the difference between the two regularizations. Basically,  $f_{[k]}$  coincides with  $f$  at those points where  $f$  has a slope not larger than  $k$ . Figure 2.2.2 illustrates the operation (2.2.3), which has the following mechanical interpretation:  $\text{gr } f_{[k]}$  is a string, which is not allowed slopes larger than  $k$ , and which is pulled upwards under the obstacle  $\text{epi } f$ .

### 2.3 Behaviour at Infinity

When studying the minimization of a function  $f \in \text{Conv } \mathbb{R}$ , the behaviour of  $f(x)$  for  $|x| \rightarrow \infty$  is crucial (assuming  $\text{dom } f = \mathbb{R}$ , the case of interest). It turns out that this behaviour is directly linked to that of the slope-function (1.1.3).

Indeed, for fixed  $x_0$ , the increasing slope-function (1.1.3) satisfies

$$\lim_{x \rightarrow \infty} s(x) = \sup_{x \neq x_0} s(x) = \sup_{x > x_0} s(x)$$

(equalities in  $\mathbb{R} \cup \{+\infty\}$ ). For  $x \rightarrow -\infty$ , its limit exists as well (in  $\mathbb{R} \cup \{-\infty\}$ ) and is likewise its infimum over  $x \neq x_0$ , or over  $x < x_0$ . To embrace the two cases in one, and to eliminate the unpleasant  $-\infty$ , it is convenient to introduce a positive variable  $t$ , playing the role of  $|x - x_0|$ : we fix a number  $d \neq 0$  and we consider

$$\lim_{t \rightarrow +\infty} \frac{f(x_0 + td) - f(x_0)}{t} = \sup_{t > 0} \frac{f(x_0 + td) - f(x_0)}{t} =: \varphi(x_0, d). \quad (2.3.1)$$

When  $d$  is positive [resp. negative],  $\varphi$  is the limiting maximal slope to the right [resp. left] of  $x_0$ . It is rather obvious that, for  $\alpha > 0$ ,  $\varphi(x_0, \alpha d) = \alpha \varphi(x_0, d)$ : in other words,  $\varphi(x_0, \cdot)$  is *positively homogeneous* (of degree 1). Hence, only the values  $\varphi(x_0, d)$  for  $d = \pm 1$  are relevant, the other values being obtained automatically.

**Theorem 2.3.1** *Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be convex. For each  $x_0 \in \mathbb{R}$  ( $= \text{dom } f$ ), the function  $\varphi(x_0, \cdot)$  of (2.3.1) is convex and does not depend on  $x_0$ .*

PROOF. The result will be confirmed in §IV.3.2, and closedness of  $\text{epi } f$  is needed, which will be proved later; nevertheless, we give the proof because it uses an interesting geometric argument. Fix  $t > 0$ ; the convexity of  $f$  implies that, for arbitrary  $d_1, d_2$  and  $\alpha \in ]0, 1[$ ,

$$f(x_0 + t\alpha d_1 + t(1 - \alpha)d_2) \leq \alpha f(x_0 + td_1) + (1 - \alpha)f(x_0 + td_2).$$

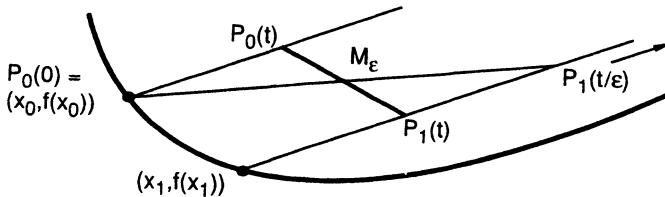
Subtract  $f(x_0)$  and divide by  $t > 0$  to see that the difference quotient  $s(x_0 + td)$  in (2.3.1) is a convex function of  $d$ . Moreover,  $\varphi(x_0, 0) = 0$ , hence Proposition 2.1.2 establishes the convexity of  $\varphi(x_0, \cdot)$ .

To show that  $\varphi(x_0, \cdot)$  does not depend on  $x_0$  is more involved. Let  $x_1 \neq x_0$  and take  $(d, r) \in \text{epi } \varphi(x_1, \cdot)$ ; we must show that  $(d, r) \in \text{epi } \varphi(x_0, \cdot)$  (then the proof will be finished, by exchanging the roles of  $x_1$  and  $x_0$ ).

By definition of  $\text{epi } \varphi(x_0, \cdot)$ , what we have to prove is that  $P_0(t) \in \text{epi } f$  (look at Fig. 2.3.1), where  $t > 0$  is arbitrary and  $P_0(t)$  has the coordinates  $x_0 + td$  and  $f(x_0) + tr$ . By definition of  $\text{epi } \varphi(x_1, \cdot)$ ,  $P_1(t) := (x_1 + td, f(x_1) + tr)$  is in  $\text{epi } f$ . Taking  $\varepsilon \in ]0, 1]$ , the key is to write the point  $M_\varepsilon$  of the picture as

$$M_\varepsilon = \varepsilon P_1(t) + (1 - \varepsilon)P_0(t) = \varepsilon P_1(t/\varepsilon) + (1 - \varepsilon)P_0(0).$$

Because  $(d, r) \in \text{epi } \varphi(x_1, \cdot)$ , the second form above implies that  $M_\varepsilon \in \text{epi } f$ ; the first form shows that, when  $\varepsilon \downarrow 0$ ,  $M_\varepsilon$  tends to  $P_0(t)$ . Admitting that  $\text{epi } f$  is closed (Theorem 3.1.1 and Proposition 3.2.2 below),  $P_0(t) \in \text{epi } f$ .  $\square$



**Fig. 2.3.1.** Inscribing a pantograph in a closed convex set

Thus, instead of  $\varphi(x_0, d)$ , the notation

$$f'_\infty(d) := \lim_{t \rightarrow +\infty} \frac{f(x_0 + td) - f(x_0)}{t} = \sup_{t > 0} \frac{f(x_0 + td) - f(x_0)}{t}$$

is more appropriate;  $x_0$  is eliminated, the symbols ' and  $\infty$  suggest that  $f'_\infty$  is a sort of "slope at infinity". This defines a new convex and positively homogeneous function, associated to  $f$ , and characterizing its behaviour at infinity in both directions  $d = \pm 1$ :

**Corollary 2.3.2** *For  $f : \mathbb{R} \rightarrow \mathbb{R}$  convex, there holds*

$$\lim_{x \rightarrow +\infty} f(x) = +\infty \iff f'_\infty(1) > 0, \quad (2.3.2)$$

$$\lim_{x \rightarrow +\infty} \frac{f(x)}{x} = +\infty \iff f'_\infty(1) = +\infty. \quad (2.3.3)$$

PROOF. By definition,  $f(x)/x \rightarrow f'_\infty(1)$  for  $x \rightarrow +\infty$ , which proves (2.3.3) and the " $\Leftarrow$ " in (2.3.2). To finish the proof, use

$$tf'_\infty(1) = f'_\infty(t) \geq f(t) - f(0) \rightarrow +\infty \text{ when } t \rightarrow +\infty$$

(remember  $0 \in \text{dom } f = \mathbb{R}$ ) and observe that  $tf'_\infty(1) \rightarrow +\infty$  certainly implies  $f'_\infty(1) > 0$ .  $\square$

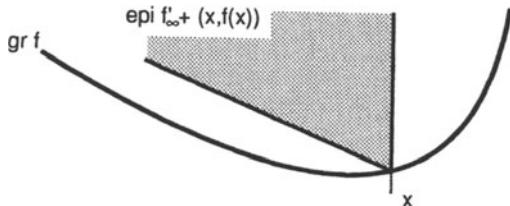
Naturally, (2.3.2) and (2.3.3) have symmetric versions, with  $x \rightarrow -\infty$ .

For (2.3.2) to hold, it suffices that  $f$  be strictly increasing on some interval of positive length. Functions satisfying (2.3.2) [resp. (2.3.3)] in both directions will be called *0-coercive* [resp. *1-coercive*]; they are important for some applications.

Geometrically, the epigraph of  $f'_\infty$  is a convex cone with apex at  $(0, 0)$ . When this apex is translated to a point  $(x, f(x))$  on  $\text{gr } f$ , the cone becomes included in  $\text{epi } f$ : in fact, the definition of  $f'_\infty$  gives for all  $y$

$$f(y) = f(x) + \frac{f(y) - f(x)}{1} \leq f(x) + f'_\infty(y - x).$$

It is then clear that  $\text{epi } f + \text{epi } f'_\infty = \text{epi } f$  (see Fig. 2.3.2), i.e. that  $f = f \downarrow f'_\infty$ . The epigraph of  $f'_\infty$  is the largest convex cone  $K$  in  $\mathbb{R}^2$  (with apex at  $(0, 0)$ ) such that  $\text{epi } f + K \subset \text{epi } f$ .

Fig. 2.3.2. The cones included in  $\text{epi } f$ 

### 3 Continuity Properties

#### 3.1 Continuity on the Interior of the Domain

Convex functions turn out to enjoy sharp continuity properties. Simple pictures suggest that a convex function may have discontinuities at the endpoints of its interval of definition  $\text{dom } f$ , but has a continuous behaviour inside. This is made precise in the following result.

**Theorem 3.1.1** *If  $f \in \text{Conv } \mathbb{R}$ , then  $f$  is continuous on  $\text{int dom } f$ . Even more: for each compact interval  $[a, b] \subset \text{int dom } f$ , there is  $L \geq 0$  such that*

$$|f(x) - f(x')| \leq L|x - x'| \quad \text{for all } x \text{ and } x' \text{ in } [a, b]. \quad (3.1.1)$$

□

Property (3.1.1) is the *Lipschitz continuity* of  $f$  on  $[a, b]$ . What Theorem 3.1.1 says is that  $f$  is *locally Lipschitzian* on the interior of its domain. It follows that the difference quotients  $[f(x) - f(x')]/(x - x')$  are themselves locally bounded, i.e. bounded on every bounded interval of  $\text{int dom } f$ .

To prove Theorem 3.1.1, the basic inequality (1.1.1) can be used on an enlargement of  $[a, b]$ , thus exhibiting an appropriate Lipschitz constant  $L$ . We prefer to postpone the proof to Remark 4.1.2 below, where another argument, coming from the differential behaviour of  $f$ , yields the Lipschitz constant directly.

It remains to see how  $f$  can behave on the boundary of its domain (assumed to be “at finite distance”). In the statement below, we recall that our notation  $x \downarrow a$  excludes the value  $x = a$ .

**Proposition 3.1.2** *Let the domain of  $f \in \text{Conv } \mathbb{R}$  have a nonempty interior and call  $a \in \mathbb{R}$  its left endpoint. Then the right-limit  $f(a_+) := \lim_{x \downarrow a} f(x)$  exists in  $\mathbb{R} \cup \{+\infty\}$ , and  $f(a) \geq f(a_+)$ .*

*Similarly, if  $b \in \mathbb{R}$  is the right endpoint of  $\text{dom } f$ , the left-limit  $f(b_-) := \lim_{x \uparrow b} f(x)$  exists in  $\mathbb{R} \cup \{+\infty\}$  and  $f(b) \geq f(b_-)$ .*

PROOF. Let  $x_0 \in \text{int dom } f$ , set  $d := -1$ ,  $t_0 := x_0 - a > 0$ . The increasing function

$$0 < t \mapsto \frac{f(x_0 + td) - f(x_0)}{t} =: q(t)$$

has a limit  $\ell \in \mathbb{R} \cup \{+\infty\}$  for  $t \uparrow t_0$ , in which case  $x_0 + td \downarrow a$ . It follows

$$f(x_0 + td) = f(x_0) + tq(t) \rightarrow f(x_0) + (x_0 - a)\ell =: f(a_+) \in \mathbb{R} \cup \{+\infty\}.$$

Then let  $t \uparrow t_0$  in the relation

$$q(t) \leq q(t_0) = \frac{f(a) - f(x_0)}{x_0 - a} \quad \text{for all } t \in ]0, t_0[$$

to obtain

$$\ell = \frac{f(a_+) - f(x_0)}{x_0 - a} \leq \frac{f(a) - f(x_0)}{x_0 - a},$$

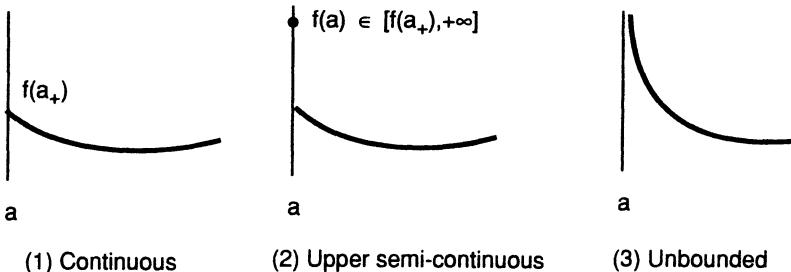
hence  $f(a_+) \leq f(a)$ . The proof for  $b$  uses the same arguments.  $\square$

The function  $q$  of the above proof is the *directional difference quotient*, already encountered in §2.3, and is nothing more than the slope-function  $[f(x) - f(x_0)]/(x - x_0)$ . We took the trouble to use it as an illustration of Remark 1.3.3, to avoid the unpleasant division by  $x - x_0 < 0$ . Furthermore it will play an important role in several dimensions.

Among other things, Proposition 3.1.2 says that  $f$  is *upper semi-continuous* (relative to  $\text{dom } f$ ) on the edge of its domain, hence on its whole domain. This property, however, is specific to the one-dimensional case, and is not true in several dimensions.

### 3.2 Lower Semi-Continuity: Closed Convex Functions

According to Proposition 3.1.2, the behaviour of a convex function at the endpoints of its domain has to resemble one of the cases illustrated on Fig. 3.2.1. We see that case (2) is somewhat “abnormal”; it is ruled out by the following definition, which thus appears as “natural”, and important for existence of solutions in minimization problems.



**Fig. 3.2.1.** Continuity properties of univariate convex functions

**Definition 3.2.1** We say that  $f \in \text{Conv } \mathbb{R}$  is *closed*, or lower semi-continuous, if

$$\liminf_{x \rightarrow x_0} f(x) \geq f(x_0) \quad \text{for all } x_0 \in \mathbb{R}. \quad (3.2.1)$$

The set of closed convex functions on  $\mathbb{R}$  is denoted by  $\overline{\text{Conv}} \mathbb{R}$ .  $\square$

Of course, (3.2.1) is an inequality in  $\mathbb{R} \cup \{+\infty\}$ . A reader not totally alert may overlook a significant detail: both  $x$  and  $x_0$  are points in the *whole of*  $\mathbb{R}$ . A closed function is therefore lower semi-continuous on the whole line, and not only relative to its domain of definition, as is the usual practice in real analysis. This is why the terminology “closed” should be preferred to lower semi-continuous.

The requirement (3.2.1) demands nothing of  $f$  beyond the closure of its domain; it is moreover true for  $x_0 \in \text{int dom } f$  (Theorem 3.1.1), the only possible problems are on the boundary of  $\text{dom } f$ . As far as the left endpoint is concerned, a closed convex function has to look like case (1) or (3) in Fig. 3.2.1 (and note: if  $a = -\infty$ , no trouble arises).

The closedness property can also be described geometrically, which by the same token justifies the terminology (note that convexity plays little role here).

**Proposition 3.2.2** *The function  $f$  is closed if and only if one of the following conditions holds:*

(i) *epi  $f$  is a closed set of  $\mathbb{R}^2$ ;*

(ii) *the sublevel-sets*

$$S_r(f) := \{x \in \mathbb{R} : f(x) \leq r\}$$

*are closed intervals of  $\mathbb{R}$  (possibly empty), for all  $r \in \mathbb{R}$ .*

□

The best way of proving this result is probably to look at Fig. 3.2.1. In practice, the closedness criterion (ii) is very useful. As an example, the function  $f'_\infty$  of §2.3 is always closed.

**Example 3.2.3** Let  $f$  be a convex function whose domain is the whole of  $\mathbb{R}$ , and let  $C$  be a nonempty closed interval. Then the “convex restriction” of  $f$  to  $C$ :

$$f_C(x) := f(x) \text{ if } x \in C, \quad +\infty \text{ otherwise}$$

is closed and convex. Its epigraph is the intersection of  $\text{epi } f$  with the vertical stripe generated by  $C$ . □

**Example 3.2.4** Let  $C$  be a nonempty interval of  $\mathbb{R}$ . The *indicator* function of  $C$  is

$$I_C(x) := \begin{cases} 0 & \text{if } x \in C, \\ +\infty & \text{otherwise.} \end{cases}$$

It is a closed convex function if and only if  $C$  is closed (its sublevel-sets are empty or  $C$ ).

The above indicator function, of constant use in convex analysis, must not be confused with the characteristic function  $\chi_C$  of measure theory, which is 1 on  $C$  and 0 outside – in fact  $\chi_C = \exp(-I_C)$ . □

Let us return to Fig. 3.2.1. In case (2) – the only bad case – we see that it is not difficult to close  $f$ : it suffices to pull  $f(a)$  down to  $f(a_+)$ . The result is in  $\overline{\text{Conv}} \mathbb{R}$ , and differs very little indeed from  $f$ .

**Definition 3.2.5** The *closure* of  $f \in \text{Conv } \mathbb{R}$  is the function defined by:

$$\text{cl } f(x) := \begin{cases} \liminf_{y \rightarrow x} f(y) & \text{if } x \in \text{cl dom } f, \\ +\infty & \text{if not.} \end{cases} \quad (3.2.2)$$

**Remark 3.2.6** The construction (3.2.2) affects  $f$  only at the endpoints of its domain. Geometrically,  $\text{cl } f$  is the function whose epigraph is the closure (in the usual topological sense) of the set  $\text{epi } f \subset \mathbb{R}^2$ . The closure of  $f$  is also the largest closed convex function minorizing  $f$ :

$$\text{cl } f(x) = \sup \{g(x) : g \in \text{Conv } \mathbb{R} \text{ and } g \leq f\}. \quad (3.2.3)$$

To close  $f$ , however, it is actually not necessary to scan the whole set of closed convex functions. It can be proved that, in (3.2.3),  $g$  can be restricted to being an *affine* function:

$$\text{cl } f(x) = \sup_{s,r} \{sx - r : sy - r \leq f(y) \text{ for all } y \in \mathbb{R}\}; \quad (3.2.4)$$

this will be established formally in Proposition IV.1.2.8. Of course, it is the *convexity* of  $f$  which allows this simplification.

The analytic operation (3.2.2), illustrated by Fig. 3.2.1, does not easily lend itself to generalizations in several dimensions (taking the lower semi-continuous hull of a function may be difficult). The operation (3.2.4) thus appears as a possible useful alternative.  $\square$

### 3.3 Properties of Closed Convex Functions

Closed convex functions are of fundamental importance in convex analysis and optimization. For one thing, the existence of a solution for the problem

$$\min \{f(x) : x \in C\}$$

requires first  $f$  to be closed (i.e. lower semi-continuous), and also  $C$  to be closed (and of course  $C \cap \text{dom } f \neq \emptyset$ ). Just as §2.1 did with convexity, it is therefore useful to know which combinations of functions preserve closedness.

**Proposition 3.3.1** Let  $f_1, \dots, f_m$  be  $m$  closed convex functions and  $t_1, \dots, t_m$  be positive numbers. If there exists  $x_0$  such that  $f_j(x_0) < +\infty$  for  $j = 1, \dots, m$ , then the function  $f := \sum_{j=1}^m t_j f_j$  is in  $\text{Conv } \mathbb{R}$ .  $\square$

Here, convexity is taken care of by Proposition 2.1.1; as for closedness, recall that *limes inferiores* are stable under addition and positive multiplication, i.e.

$$\liminf t(u_k + v_k) \geq t \liminf u_k + t \liminf v_k$$

(an inequality in  $\mathbb{R} \cup \{+\infty\}$ ).

The next result also comes easily, since an intersection of closed sets is closed.

**Proposition 3.3.2** Let  $\{f_j\}_{j \in J}$  be a family of closed convex functions. If there exists  $x_0 \in \mathbb{R}$  such that  $\sup_{j \in J} f_j(x_0) < +\infty$ , then the function  $f := \sup_{j \in J} f_j$  is in  $\text{Conv } \mathbb{R}$ .  $\square$

**Example 3.3.3** Let  $f : \mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$  be a function not identically  $+\infty$  (but not necessarily convex) minorized by some affine function: there are  $s_0, r_0$  such that  $f(x) \geq s_0x - r_0$  for all  $x$ . Then the so-called *conjugate* function of  $f$ :

$$\mathbb{R} \ni s \mapsto \sup \{sx - f(x) : x \in \text{dom } f\}$$

is finite at least at  $s_0$ , and it is closed convex (as a supremum of affine functions!).  $\square$

The case of the infimal convolution is much more delicate (it is an infimum, hence a limit, and the inequality signs go the wrong direction, in contrast with Proposition 3.3.2).

**Remark 3.3.4** It can indeed be proved that the inf-convolution of two functions of  $\overline{\text{Conv}} \mathbb{R}$  is still closed, but this is a specific result of the univariate case: its extension to several variables requires some additional assumption.

To accept the closedness of a one-dimensional inf-convolution, a first key is to realize that, if  $a$  is the left endpoint of the domain of  $f = f_1 \downarrow f_2$ , then  $a = a_1 + a_2$  and  $f(a) = f_1(a_1) + f_2(a_2)$  where, for  $i = 1, 2$ ,  $a_i$  is the left endpoint of  $\text{dom } f_i$ . Thus, for  $k = 1, 2, \dots$ , take  $x_1^k$  and  $x_2^k$  satisfying:  $x_i^k \in \text{dom } f_i$  and  $x_1^k + x_2^k = x_k \downarrow a$ ; a second key is then to see that  $x_i^k \downarrow a_i$  for  $i = 1, 2$ . If, in addition,

$$f_1(x_1^k) + f_2(x_2^k) \leq (f_1 \downarrow f_2)(x_k) + 1/k,$$

it suffices to pass to the limit, using the properties  $f(x_i^k) \rightarrow f(a_i)$  for  $i = 1, 2$ .  $\square$

Finally, the case of limit-functions of §2.2 is of course hopeless. The traditional example  $x \mapsto f_k(x) := |x|^k$  converges pointwise when  $k \rightarrow +\infty$  to

$$x \mapsto f(x) = \begin{cases} 0 & \text{if } x \in ]-1, +1[, \\ 1 & \text{if } x \in \{-1, +1\}, \\ +\infty & \text{otherwise,} \end{cases}$$

which is not closed. Some “uniformity” in the convergence is required, and this establishes a link between Remark 2.2.2 and Theorem 3.1.1.

## 4 First-Order Differentiation

Monotonicity of the slope-function (1.1.3) provides convex functions with rather astonishing properties of “one-sided differentiability”, which allow the introduction of a substitute for the concept of derivative: the “set of subderivatives” of a convex function at a point of its domain. A “subdifferential calculus” can then be developed for convex functions, which plays the role of differential calculus in the  $C^1$  case, and which gives similar results.

## 4.1 One-Sided Differentiability of Convex Functions

**Theorem 4.1.1** Let  $f \in \text{Conv } \mathbb{R}$ . At all  $x_0$  in the interior of its domain,  $f$  admits a finite left-derivative and a finite right-derivative:

$$D_{-}f(x_0) := \lim_{x \uparrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = \sup_{x < x_0} \frac{f(x) - f(x_0)}{x - x_0} \quad (4.1.1)$$

$$D_{+}f(x_0) := \lim_{x \downarrow x_0} \frac{f(x) - f(x_0)}{x - x_0} = \inf_{x > x_0} \frac{f(x) - f(x_0)}{x - x_0}. \quad (4.1.2)$$

They satisfy

$$D_{-}f(x_0) \leq D_{+}f(x_0). \quad (4.1.3)$$

PROOF. Apply the criterion 1.1.4 of increasing slopes: the difference quotient involved in (4.1.1), (4.1.2) is just the slope-function  $s$ . For any two points  $x, x'$  in  $\text{int dom } f$  satisfying  $x < x_0 < x'$ ,  $s(x)$  and  $s(x')$  are finite numbers satisfying  $s(x) \leq s(x')$ . Furthermore, when  $x \uparrow x_0$  [resp.  $x \downarrow x_0$ ],  $s(x)$  increases [resp.  $s(x')$  decreases], hence they both converge, say as described by the notation (4.1.1), (4.1.2); this proves (4.1.3) at the same time.  $\square$

**Remark 4.1.2** Proof of Theorem 3.1.1. Take  $[a, b] \subset \text{int dom } f$  ( $a < b$ : there is nothing to prove if  $a = b$ ) and  $a \leq x < x' \leq b$ ; if  $a < x$ , use (4.1.1) and (4.1.2) written at appropriate points to obtain

$$\begin{aligned} D_{+}f(a) &\leq \frac{f(x) - f(a)}{x - a} \leq D_{-}f(x) \leq D_{+}f(x) \leq \\ &\leq \frac{f(x') - f(x)}{x' - x} \leq D_{-}f(x') \leq \text{etc.} \leq D_{-}f(b); \end{aligned}$$

note that the relevant inequalities hold as well if  $x = a$ . This proves (3.1.1) with  $L = \max\{-D_{+}f(a), D_{-}f(b)\}$ .  $\square$

A sort of differentiability being thus established on the interior of  $\text{dom } f$ , what can be said about its endpoints? Let again  $a$  be its left endpoint, as in Fig. 3.2.1. First of all, the whole concept is meaningless if  $a \notin \text{dom } f$  (case 3), and the very definition shows that  $D_{-}f(a) = -\infty$ . As for the right-derivative, its existence is ruled out if  $f$  is not closed (case 2); finally, the criterion of increasing slopes tells us that  $D_{+}f(a)$  does exist, but in  $\mathbb{R} \cup \{-\infty\}$ . Let us summarize these observations:

**Proposition 4.1.3** For  $x_0$  on the left [resp. right] endpoint of  $\text{dom } f$ , (4.1.2) [resp. (4.1.1)] holds as an equality in  $\mathbb{R} \cup \{-\infty\}$  [resp.  $\mathbb{R} \cup \{+\infty\}$ ].  $\square$

**Remark 4.1.4** Our notations deserve comment, since the left- and right-derivatives are usually denoted by  $f'_-$  and  $f'_+$  in real analysis.

In Sections 2.3 and 3.1, we have encountered the directional difference quotient

$$q_{x_0, d}(t) := q(t) := \frac{f(x_0 + td) - f(x_0)}{t}$$

at  $x_0 \in \text{dom } f$  in the direction  $d \in \mathbb{R}$ . It is monotonic, satisfies  $q_{x_0, \alpha d}(t) = q_{x_0, d}(\alpha t)$  for  $\alpha > 0$ , its supremum over  $t > 0$  is  $f'_\infty(d)$ , its infimum over  $t > 0$  is its limit for  $t \downarrow 0$ . Of central importance in convex analysis is the corresponding *directional derivative* of  $f$  at  $x_0$  in the direction  $d$  ( $= \pm 1$ ), traditionally denoted by:

$$f'(x_0, d) := \lim_{t \downarrow 0} \frac{f(x_0 + td) - f(x_0)}{t} = \inf_{t > 0} \frac{f(x_0 + td) - f(x_0)}{t}. \quad (4.1.4)$$

Now the concept of derivative in real analysis contains several distinct objects.

- From its definition, it is first a *number*, say  $Df(x_0)$ , computed as a limit.
- On the other hand, this number is also a *linear form* which, when applied to a  $dx$ , yields the corresponding  $df$ :

$$df = Df(x_0) \cdot dx.$$

- It is therefore also the *value* of this linear form at  $dx = 1$  (naturally, the fact that this linear form happens to depend on  $x_0$  does not help to clarify the matter).

The notation (4.1.4) definitely represents this last interpretation, namely the value of the linear form at  $d$ ; by contrast, (4.1.1), (4.1.2) is the linear form itself (or a pair of such) – hence our choice of two definitely different notations. As a matter of fact, the two things do not coincide:

$$f'(x_0, 1) = D_+ f(x_0) \quad \text{but} \quad D_- f(x_0) = -f'(x_0, -1) \quad (4.1.5)$$

(beware of the minus sign!).

For this reason, we will generally denote by  $Df(x)$  the ordinary derivative of a function  $f$  differentiable at  $x$ . It is only when no confusion is possible (mainly in subsequent chapters) that we will use the more classical notation  $f'(x)$ , and also  $f''(x)$  for the second derivative.  $\square$

We are now in a position to introduce the notion of subderivative of a convex function.

**Definition 4.1.5** Let  $f \in \text{Conv } \mathbb{R}$ . We say that  $s \in \mathbb{R}$  is a *subderivative* of  $f$  at  $x \in \text{dom } f$  when

$$D_- f(x) \leq s \leq D_+ f(x). \quad (4.1.6)$$

The *subdifferential*  $\partial f(x)$  is the set of all subderivatives of  $f$  at  $x$ . It is the line-segment  $[D_- f(x), D_+ f(x)]$  when  $D_- f(x)$  and  $D_+ f(x)$  are finite.  $\square$

Thus, it is clear that

- for  $x_0 \in \text{int dom } f$ , the subdifferential  $\partial f(x_0)$  is a nonempty compact interval: this results from Theorem 4.1.1;
- for  $x_0 \notin \text{dom } f$ ,  $\partial f(x_0)$  is empty;
- at an endpoint point such as  $a$  of Fig. 3.2.1,  $\partial f$  is certainly empty if  $f$  is not closed; if  $f$  is closed,  $\partial f$  may be empty (case of a vertical slope), otherwise it has the form  $]-\infty, D_+ f(a)]$ .

In the language of Remark 4.1.4, a subderivative suggests a linear form. It can also be characterized in terms of the values of this linear form, namely:  $s$  is a subderivative of  $f$  at  $x_0$  if and only if

$$f(x) \geq f(x_0) + s(x - x_0) \quad \text{for all } x \in \mathbb{R}, \quad (4.1.7)$$

a result which comes directly from the property of increasing slopes (to prove it, avoid division by  $x - x_0$ , but set  $x - x_0 = td$  with  $d = \pm 1$ , and divide by  $t > 0$ ). The linear form attached to a subderivative thus defines an affine function which minorizes  $f$ . Incidentally, (4.1.7) readily gives a necessary and sufficient condition for  $x_0$  to minimize  $f$ , namely:

$$x_0 \text{ minimizes } f \in \text{Conv } \mathbb{R} \iff 0 \in \partial f(x_0). \quad (4.1.8)$$

**Remark 4.1.6** Comparing (4.1.5) with (4.1.6), we can see that the subdifferential is also characterized as

$$\partial f(x_0) = \{s \in \mathbb{R} : sd \leq f'(x_0, d) \text{ for all } d \in \mathbb{R}\};$$

or alternatively, the directional derivative is characterized (whenever  $\partial f(x) \neq \emptyset$ ) as

$$f'(x_0, d) = \sup \{sd : s \in \partial f(x)\}.$$

The directional derivative takes its values in  $\mathbb{R} \cup \{\pm\infty\}$ . All these observations concerning the directional derivative will have their importance when going to several dimensions.  $\square$

Altogether, the subdifferential defines a *multipunction* from  $\mathbb{R}$  (or rather  $\text{dom } f$ ) to the subsets of  $\mathbb{R}$ . See Fig. 4.1.1 for a possible behaviour of this multifunction; and see §A.5 for an introduction to the most important concepts of set-valued analysis.

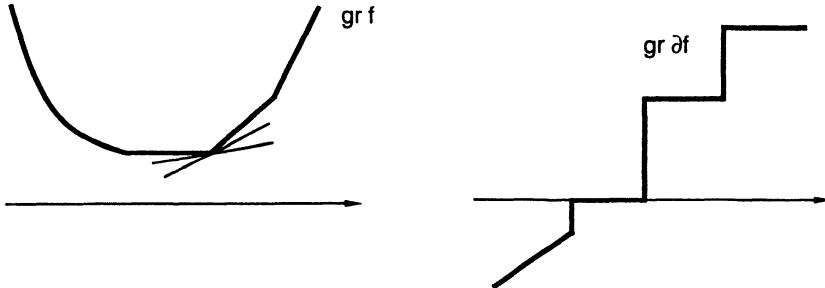


Fig. 4.1.1. A typical subdifferential mapping

**Remark 4.1.7** Denoting by  $\text{dom } \partial f$  the domain of  $\partial f$ , it is a consequence of Theorem 4.1.1 and of the definition (4.1.6) that

$$\text{dom } f \supset \text{dom } \partial f \supset \text{int dom } f.$$

Remembering (4.1.7), we see that a convex function is minorized by some affine function whenever the interior of its domain is nonempty: just take a subderivative. Now, the only convex functions whose domain has an empty interior are very special: up to a constant, they are indicators of one single point, say  $x_0$ . For them also, the existence of a minorizing affine function is clear (actually,  $\partial f(x_0) = \mathbb{R}$  in this case). We conclude that any convex function is minorized by some *affine* function.  $\square$

A *kink* (or corner-point) is a point  $x$  where  $\partial f(x)$  is not a singleton. If  $x$  is not a kink, then  $f$  is differentiable at  $x$ :  $Df(x) = D_- f(x) = D_+ f(x)$ .

**Example 4.1.8** Let  $a_1 < a_2 < \dots < a_m$  be  $m$  real numbers and  $t_1, \dots, t_m$  be positive. We consider  $f$  defined on  $\mathbb{R}$  by

$$f(x) := \sum_{j=1}^m t_j |x - a_j|.$$

Convexity and closedness of  $f$ , if not considered as trivial, come from Propositions 3.3.1 and 3.3.2 (recall that  $|z| = \max\{z, -z\}$ ). This  $f$  is differentiable at all  $x$  except at  $a_1, \dots, a_m$  and there holds

$$\partial f(x) = \begin{cases} \sum_{\{j: a_j < x\}} t_j - \sum_{\{j: a_j > x\}} t_j & \text{if } x \notin \{a_1, \dots, a_m\} \\ \sum_{\{j: a_j < x\}} t_j - \sum_{\{j: a_j > x\}} t_j + [-t_{j_0}, t_{j_0}] & \text{if } x = a_{j_0}. \end{cases}$$

If we want to minimize  $f$ , an algorithm can then be conceived, based on scanning the interval  $[a_1, a_m]$  from left to right. The algorithm stops when the minimality condition (4.1.8) is met.  $\square$

**Example 4.1.9** Consider the function  $x \mapsto f(x) := \int_0^x \varphi(u) du$ , where  $\varphi(u) = u$  for  $u \leq 0$  and, for  $u > 0$ ,  $\varphi$  is defined via the integer part of  $1/u$ :

$$\varphi(u) = \frac{1}{1+k} \quad \text{if } k+1 > \frac{1}{u} \geq k \in \mathbb{N}$$

( $\varphi$  oscillates between the functions  $x$  and  $x/(1+x)$ , see Fig. 4.1.2). The function  $f$  is convex (criterion 1.1.4 of increasing slopes), has 0-derivative at 0, but is differentiable on no segment  $]0, u[$ : its kinks are at  $1/k$ , with  $\partial f(1/k) = [1/(k+1), 1/k]$ ,  $k = 1, 2, \dots$ .  $\square$

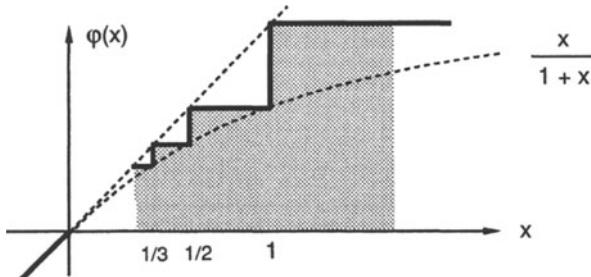


Fig. 4.1.2. Infinitely many discontinuities

## 4.2 Basic Properties of Subderivatives

Convexity implies that  $f$  is differentiable at “many” points, and that  $\partial f(x)$  behaves very nicely when  $x$  varies:

**Theorem 4.2.1** For  $f \in \text{Conv } \mathbb{R}$ , the following properties hold:

(i) *The multifunction  $\partial f$  is increasing on its domain, in the sense that*

$$s_1 \leq s_2 \text{ whenever } s_1 \in \partial f(x_1), s_2 \in \partial f(x_2) \text{ and } x_1 < x_2. \quad (4.2.1)$$

(ii) *The set of points where  $f$  fails to be differentiable is at most countable.*

(iii) *For all  $x_0 \in \text{int dom } f$ , the sets  $\partial f(x)$  converge increasingly to  $D_- f(x_0)$  when  $x \uparrow x_0$ , in the following sense: for all  $\varepsilon > 0$ , there is  $\delta > 0$  such that*

$$s \in \partial f(x) \text{ and } x \in ]x_0 - \delta, x_0[ \implies s \in [D_- f(x_0) - \varepsilon, D_- f(x_0)]; \quad (4.2.2)$$

*likewise,  $\partial f(x)$  converges decreasingly to  $D_+ f(x_0)$  when  $x \downarrow x_0$  (symmetric definition).*

PROOF. Start from (4.1.1), (4.1.2) and Proposition 4.1.3: for any two points  $x_1 < x_2$  in  $\text{dom } \partial f \subset \text{dom } f$ ,

$$s_1 \leq D_+ f(x_1) \leq \frac{f(x_2) - f(x_1)}{x_2 - x_1} \leq D_- f(x_2) \leq s_2 \quad (4.2.3)$$

whenever  $s_i \in \partial f(x_i)$ ,  $i = 1, 2$  (see Definition 4.1.5). This proves (i).

It follows from (4.2.3) that the intervals  $\partial f(x_1)$  and  $\partial f(x_2)$  are disjoint if  $x_1 \neq x_2$ . Let  $\Delta$  be the set of points in  $\text{int dom } f$  where  $f$  fails to be differentiable, i.e. where  $D_- f(x) < D_+ f(x)$ . Then  $\{]D_- f(x), D_+ f(x)[\}_{x \in \Delta}$  form a collection of nonempty disjoint intervals of  $\mathbb{R}$ ; this collection is therefore at most countable, and so is  $\Delta$ .

Now let  $x_0 \in \text{int dom } f$ . In view of (i),

$$\limsup \{s : s \in \partial f(x), x \uparrow x_0\} \leq D_- f(x_0). \quad (4.2.4)$$

Take  $\text{dom } f \ni x < x' < x_0$  and write (4.1.1) with  $x_0$  replaced by  $x'$ : by definition of  $\partial f$ ,

$$\frac{f(x) - f(x')}{x - x'} \leq s' \quad \text{for all } s' \in \partial f(x').$$

Letting  $x' \uparrow x_0$  and using the continuity of  $f$  at  $x_0$ :

$$\frac{f(x) - f(x_0)}{x - x_0} \leq \liminf \{s' : s' \in \partial f(x'), x' \uparrow x_0\}.$$

It remains to let  $x \uparrow x_0$  and to compare with (4.2.4) to obtain (4.2.2).  $\square$

Property (iii) means that, when  $x$  tends to  $x_0$  while staying on the same side of  $x_0$ , the whole set  $\partial f(x)$  shrinks to one single particular endpoint of  $\partial f(x_0)$ , namely the half-derivative corresponding to the side that  $x$  comes from. The proof can of course be extended when  $x_0$  is the right endpoint of  $\text{dom } f$ , provided that  $f$  is closed (the continuity of  $f$  is explicitly used). Finally, remember Example 4.1.9, which shows that existence of a (half-)derivative of  $f$  does not imply single-valuedness of  $\partial f$  in a neighborhood.

**Remark 4.2.2** What Theorem 4.2.1(iii) says is that  $D_- f$  is left-continuous and  $D_+ f$  is right-continuous, wherever they exist. This double result can be condensed into one with the help of the notation (4.1.4): for all  $x \in \text{dom } f$  and all  $d$ , the following holds in  $\mathbb{R} \cup \{\pm\infty\}$

$$f'(x + td, d) \downarrow f'(x, d) \quad \text{when } t \downarrow 0.$$

□

Needless to say,  $\partial f(x)$  is the ordinary derivative  $\{Df(x)\}$  whenever  $f$  is differentiable at  $x$  – and this occurs for “most”  $x$ . More can be said about this case:

**Corollary 4.2.3** *With  $f \in \text{Conv } \mathbb{R}$ , let  $x_0 \in \text{int dom } f$  be a point where  $f$  is differentiable, with derivative  $Df(x_0) = D_- f(x_0) = D_+ f(x_0)$ . Then  $\partial f(x)$  converges to  $Df(x_0)$  when  $x \rightarrow x_0$ .*

*In particular,  $f$  is continuously differentiable at  $x_0$  whenever it is differentiable in a neighborhood of  $x_0$ .* □

As shown for example by (4.2.3), the difference quotient between  $x_1$  and  $x_2$  lies between the slopes at  $x_1$  and  $x_2$ . This suggests that some mean-value theorem should hold as an equality. Such is indeed the case:

**Theorem 4.2.4 (Mean-Value Theorem)** *Let  $f \in \text{Conv } \mathbb{R}$  and let  $[a, b] \subset \text{dom } f$  with  $a < b$ . Then there exists  $c \in ]a, b[$  such that*

$$\frac{f(b) - f(a)}{b - a} \in \partial f(c). \quad (4.2.5)$$

PROOF. As usual in this context, consider the auxiliary function

$$g(x) := f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a).$$

It is continuous on  $[a, b]$ , it has been constructed so that  $g(a) = g(b) = 0$ , so it is minimal at some  $c \in ]a, b[$ . Also, inspection of the left- and right-derivatives shows that

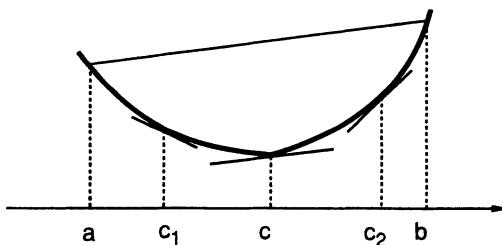
$$\partial g(c) = \partial f(c) - \left\{ \frac{f(b) - f(a)}{b - a} \right\}.$$

Thus, the minimality condition (4.1.8) characterizing  $c$  is exactly (4.2.5). □

**Remark 4.2.5** The particular subderivative singled out by (4.2.5) is a convex combination of  $D_- f(c)$  and  $D_+ f(c)$ . Now Fig. 4.2.1 suggests the following construction: take arbitrary  $s_1 \in \partial f(c_1)$  and  $s_2 \in \partial f(c_2)$  with  $a \leq c_1 < c < c_2 \leq b$ . From the monotonicity property (4.2.1),  $\partial f(c) \subset [s_1, s_2]$ , hence (4.2.5) can also be represented as a convex combination of these  $s_1$  and  $s_2$ .

Another observation is that the locally Lipschitzian  $f$  is, at least locally, absolutely continuous (see §A.6). Indeed, if  $s : \text{dom } \partial f \rightarrow \mathbb{R}$  is an arbitrary selection  $s(u) \in \partial f(u)$ , there holds

$$f(b) - f(a) = \int_a^b s(u)du. \quad (4.2.6)$$



**Fig. 4.2.1.** Mean-value theorem for convex functions

The symbolic writing

$$f(x) - f(a) = \int_a^x \partial f(u) du \quad \text{for all } x \in [a, b]$$

gives one more mean-value theorem. An easy consequence of this integral representation is: if  $f$  and  $g$  are closed convex functions such that

$$\partial f(x) \subset \partial g(x) \quad \text{for all } x \in [a, b] \subset \text{int dom } f \cap \text{int dom } g,$$

then  $f$  and  $g$  differ by a constant on  $[a, b]$ .  $\square$

### 4.3 Calculus Rules

When convex functions are combined so as to form a new convex function, their subdifferentials obey calculus rules resembling those of ordinary differential calculus. A difference, however, is that there are operations preserving convexity which do not preserve differentiability (like taking a pointwise maximum). Indeed, computing the subdifferential of a composite function  $f$  amounts to computing the half-derivatives  $D_- f$  and  $D_+ f$ ; difficulties might occur, however, at the endpoints of the various domains involved, where these half-derivatives might take on infinite values.

**Proposition 4.3.1** *Let  $f_1, \dots, f_m$  be  $m$  convex functions, all finite in the neighborhood of some point  $x$ , and let  $t_1, \dots, t_m$  be positive numbers. Then, for  $f := \sum_{j=1}^m t_j f_j$ ,*

$$\partial f(x) = \sum_{j=1}^m t_j \partial f_j(x).$$

PROOF. Just apply to the half-derivatives  $D_- f_j$  and  $D_+ f_j$  the standard calculus on limits, and use the addition of compact intervals of  $\mathbb{R}$ .  $\square$

**Proposition 4.3.2** *Let  $f_1, \dots, f_m$  and  $x$  be as described in Proposition 4.3.1. Setting  $f := \max_j f_j$ , let*

$$J(x) := \{j = 1, \dots, m : f_j(x) = f(x)\}$$

*be the set of active indices at  $x$ . Then  $\partial f(x)$  is the smallest interval containing each  $\partial f_j(x)$ ,  $j \in J(x)$ .*

PROOF. Observe by direct calculation that

$$D_+ f(x) = \max_{j \in J(x)} D_+ f_j(x) \quad \text{and} \quad D_- f(x) = \min_{j \in J(x)} D_- f_j(x). \quad \square$$

As an illustration, suppose that all the  $f_j$ 's are differentiable, with derivatives  $Df_j(x)$ . Then

$$\partial f(x) = [\min_{j \in J(x)} Df_j(x), \max_{j \in J(x)} Df_j(x)].$$

If, in addition, there is only one active index at  $x$ , say  $j(x)$  (a situation likely to happen most of the time, draw a picture), then  $f$  is differentiable at  $x$  and  $Df(x) = Df_{j(x)}(x)$ .

Note: Propositions 4.3.1 and 4.3.2 could be extended to an  $x$  on the boundary of some  $\text{dom } f_j$ , provided that all the  $\partial f_j(x)$ 's are nonempty (then, calculus on the half-derivatives is to be understood in  $\mathbb{R} \cup \{\pm\infty\}$ ).

**Proposition 4.3.3** *With  $f_1$  and  $f_2$  convex and minorized by a common affine function, let  $f = f_1 \downarrow f_2 \in \text{Conv } \mathbb{R}$ . Take  $x \in \text{dom } f = \text{dom } f_1 + \text{dom } f_2$ , and suppose that there exists  $(x_1, x_2) \in \text{dom } f_1 \times \text{dom } f_2$  such that the infimal convolution is exact at  $x = x_1 + x_2$ , i.e.  $f(x) = f_1(x_1) + f_2(x_2)$ . Then*

$$\partial f(x) = \partial f_1(x_1) \cap \partial f_2(x_2).$$

PROOF. Use (4.1.7) and decompose any  $y \in \mathbb{R}$  as  $y = y_1 + y_2$ : a slope  $s$  belongs to  $\partial f(x)$  if and only if, for all  $(y_1, y_2) \in \mathbb{R}^2$ ,

$$f_1(y_1) + f_2(y_2) \geq f_1(x_1) + f_2(x_2) + s(y_1 + y_2 - x). \quad (4.3.1)$$

Setting successively  $y_i = x_i$ ,  $i = 1, 2$ , we see that  $s \in \partial f(x_i)$ ,  $i = 1, 2$ .

Conversely, if  $s \in \partial f(x_i)$  for  $i = 1, 2$ , we have

$$f_i(y_i) \geq f_i(x_i) + s(y_i - x_i) \quad \text{for } i = 1, 2 \text{ and all } y_i \in \mathbb{R};$$

then (4.3.1) follows by mere addition.  $\square$

It is worth noting in the above result that, if either  $f_1$  or  $f_2$  is differentiable, so is  $f_1 \downarrow f_2$ . A particularly interesting application is the regularization of Moreau-Yosida (2.2.2): let  $f \in \text{Conv } \mathbb{R}$  and  $y_{(k)}(x)$  be the unique solution of (2.2.2); since the function  $x \mapsto \frac{1}{2}kx^2$  has the derivative  $kx$ , we conclude that  $f_{(k)}$  is differentiable, and that  $Df_{(k)}(x) = k[y_{(k)}(x) - x]$ . See again Example 1.1.5 and Fig. 2.2.1.

**Proposition 4.3.4** *Let  $\{f_k\}_{k \in \mathbb{N}}$  be a sequence of convex functions converging pointwise to a (convex) function  $f$  and take  $x \in \text{dom } f$  (assumed nonempty). For any sequence  $s_k \in \partial f_k(x)$ , the cluster points of  $\{s_k\}$  are all in  $\partial f(x)$ .*  $\square$

With set-theoretic notation, the property expressed in this result can be written as

$$\lim_{k \rightarrow \infty} \text{ext } \partial f_k(x) \subset \partial f(x); \quad (4.3.2)$$

(see §A.5: the *limes exterior* is the set of all cluster-points). Just as with Proposition 4.3.3, the proof uses the characterization (4.1.7): it suffices to pass to the limit in

$$f_k(y) \geq f_k(x) + s_k(y - x) \quad \text{for all } y \in \mathbb{R}$$

(a technical point is that, since the limit  $f$  is finite at  $x$  by assumption, then necessarily  $f_k(x)$  is also finite for  $k$  large enough).

Counter-examples to the converse inclusion in (4.3.2) are known even in classical differential calculus, for instance  $x \mapsto f_k(x) = \sqrt{x^2 + 1/k}$ : when  $k \rightarrow +\infty$ ,  $f_k$  converges (even uniformly) to  $|x|$  and

$$Df_k(0) \equiv 0 \rightarrow 0 \in [-1, +1] = \partial(|\cdot|)(0).$$

We conclude this section with a simple example: taking  $f$  to be a convex function and  $C$  a closed interval included in  $\text{int dom } f$ , consider the minimization problem

$$\inf \{f(x) : x \in C\}. \quad (4.3.3)$$

With the help of the indicator function of Example 3.2.4, it can be transformed to the obviously equivalent problem:

$$\inf \{g(x) : x \in \mathbb{R}\}, \quad \text{where } g = f + I_C,$$

in which the constraint is hidden; formally, it suffices to study unconstrained problems. Furthermore, (4.1.8) tells us that this in turn is equivalent to finding  $x$  such that  $0 \in \partial g(x)$ , which can be further expressed as:

$$-\infty \leq D_- g(x) \leq 0 \leq D_+ g(x) \leq +\infty,$$

or also, in terms of the directional derivative:

$$g'(x, d) \geq 0 \quad \text{for all } d \text{ or for } d = \pm 1.$$

Existence of such a solution is linked to the behaviour of  $g(x)$  when  $|x| \rightarrow \infty$ , see §2.3. We just mention a result emerging from the continuity properties of the half-derivatives: if there exist  $x_1$  and  $x_2$  with  $x_1 \leq x_2$  and  $D_+ g(x_1) \geq 0$ ,  $D_- g(x_2) \leq 0$ , then there exists a solution in  $[x_1, x_2]$ .

Our assumption  $C \subset \text{int dom } f$  enables the use of Proposition 4.3.1: the subdifferential  $\partial I_C(x)$  is clearly empty for  $x \notin C$ ,  $\{0\}$  for  $x \in \text{int } C$ , and  $]-\infty, 0]$  (resp.  $[0, +\infty[$ ) for  $x$  on the left (resp. right) endpoint of  $C$ . It is then easy to characterize an optimal solution:  $x$  solves (4.3.3) if and only if it satisfies one of the three properties:

- either  $x \in \text{int } C$  and  $0 \in \partial f(x)$ ;
- or  $x$  is the left endpoint of  $C$  and  $D_+ f(x) \geq 0$ ;
- or  $x$  is the right endpoint of  $C$  and  $D_- f(x) \leq 0$ .

## 5 Second-Order Differentiation

First-order differentiation of a convex function  $f$  results in the increasing derivatives  $D_- f(\cdot)$  and  $D_+ f(\cdot)$  – or, in a condensed way, in the increasing multifunction  $\partial f$ . In view of Lebesgue's differentiation theorem (§A.6), a convex function is therefore “twice differentiable almost everywhere”, giving way to some sort of second derivatives. The behaviour of such second derivatives, however, is much less pleasant than that of first derivatives. In a word, anything can happen: they can oscillate, or approach infinity anywhere; their only certain property is nonnegativity.

## 5.1 The Second Derivative of a Convex Function

First of all, we specify what we mean by “twice differentiability” for a convex function.

**Definition 5.1.1** Let  $f \in \text{Conv } \mathbb{R}$ . We say that the multifunction  $\partial f$  is differentiable at  $x \in \text{int dom } f$  when

- (i)  $\partial f(x)$  is a singleton  $\{Df(x)\}$  (which is thus the usual derivative of  $f$  at  $x$ ), and
- (ii) there is a real number  $D_2 f(x)$  such that

$$\lim_{h \rightarrow 0} \frac{\partial f(x+h) - Df(x)}{h} = \{D_2 f(x)\}, \quad (5.1.1)$$

i.e.:  $\forall \varepsilon > 0, \exists \delta > 0$  such that  $|h| \leq \delta$  and  $s \in \partial f(x+h)$  implies

$$|s - Df(x) - D_2 f(x)h| \leq \varepsilon |h|. \quad (5.1.2)$$

□

Putting  $s$  on each endpoint of  $\partial f(x+h)$  in (5.1.2), one sees that differentiability of  $\partial f$  implies the usual differentiability of  $D_- f$  and  $D_+ f$  at  $x$ . Conversely, it is not too difficult to see via Theorem 4.2.1 that differentiability of  $D_- f$  implies differentiability of  $\partial f$  at  $x$ , and of  $D_+ f$  as well. In a word: differentiability at  $x$  of the multifunction  $\partial f$ , or of  $D_- f$ , or of  $D_+ f$ , are three equivalent properties.

Note that this differentiability does not force  $\partial f$  to be single-valued in a neighborhood of  $x$ : indeed, the  $\partial f$  of Example 4.1.9 is differentiable at 0, with  $D_2 f(0) = 1$ . Geometrically, the multifunction  $\partial f$  is differentiable when it is as displayed in Fig. 5.1.1: all the possible curves  $h \mapsto s(h) \in \partial f(x+h)$  have the same tangent, of equation  $s(h) = Df(x) + D_2 f(x)h$ .

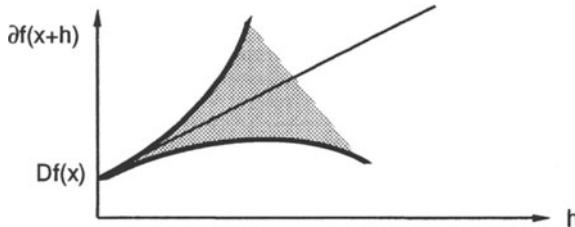


Fig. 5.1.1. Allowed values for a differentiable multifunction

In real analysis, a function  $f$  has a second derivative  $\ell$  at  $x$  if

$$\frac{Df(x+h) - Df(x)}{h} \quad \text{has a limit } \ell \text{ for } h \rightarrow 0; \quad (5.1.3)$$

this means that  $Df$  has a first-order development near  $x$ :

$$Df(x+h) = Df(x) + \ell h + o(|h|).$$

Then  $f$  itself has a second-order development near  $x$ :

$$f(x+h) = f(x) + Df(x)h + \frac{1}{2}\ell h^2 + o(h^2). \quad (5.1.4)$$

Conversely, it is generally not true that a second-order development of  $f$  implies the existence of a second derivative. In the convex case, however, equivalence is obtained if the differentiability definition (5.1.1) is used as a substitute for (5.1.3):

**Theorem 5.1.2** *Let  $f \in \text{Conv } \mathbb{R}$  and  $x \in \text{int dom } f$ . Then the two statements below are equivalent:*

- (i)  $\partial f$  is differentiable at  $x$  in the sense of (5.1.1);
- (ii)  $f$  has a second-order development (5.1.4) at  $x$  with  $\ell = D_2 f(x)$ .

PROOF. [(i)  $\Rightarrow$  (ii)] Given  $\varepsilon > 0$ , take  $|h|$  so small that, for all  $|u| \leq |h|$ ,

$$-\varepsilon|u| \leq s(x+u) - Df(x) - D_2 f(x)u \leq \varepsilon|u|.$$

Integrate from 0 to  $h$  to obtain with (4.2.6):

$$|f(x+h) - f(x) - Df(x)h - \frac{1}{2}D_2 f(x)h^2| \leq \frac{1}{2}\varepsilon h^2.$$

[(ii)  $\Rightarrow$  (i)] Fix  $\theta$  arbitrarily in  $]0, 1[$ ; develop  $f(x+h)$  and  $f(x+\theta h)$  according to (5.1.4) and obtain by subtraction

$$f(x+h) - f(x+\theta h) = (1-\theta)Df(x)h + \frac{1}{2}\ell(1-\theta^2)h^2 + o(h^2).$$

From the mean-value theorem 4.2.4, there is  $c$  between  $x+h$  and  $x+\theta h$ , and  $s \in \partial f(c)$  such that

$$s = \frac{f(x+h) - f(x+\theta h)}{(1-\theta)h}.$$

Applying the definition (5.1.4) to  $f(x+h)$  and  $f(x+\theta h)$ , we therefore get

$$s = Df(x) + \frac{1}{2}\ell(1+\theta)h + o(h).$$

Now apply the monotonicity property (4.2.1): assuming for example  $h > 0$ ,

$$\partial f(x+\theta h) \leq s \leq \partial f(x+h) \quad (5.1.5)$$

so that we obtain

$$\begin{aligned} \frac{\partial f(x+\theta h) - Df(x)}{\theta h} &\leq \frac{s - Df(x)}{\theta h} = \frac{1}{2}\ell \frac{1+\theta}{\theta} + \frac{o(h)}{h} \\ \frac{\partial f(x+h) - Df(x)}{h} &\geq \frac{s - Df(x)}{h} = \frac{1}{2}\ell(1+\theta) + \frac{o(h)}{h}. \end{aligned}$$

If  $h < 0$ , inequalities are reversed in (5.1.5) but the division by  $h$  reproduces the same last two inequalities.

Finally, let  $h \rightarrow 0$  ( $\theta$  is still fixed):

$$\begin{aligned} \limsup_{h \rightarrow 0} \frac{\partial f(x+h) - Df(x)}{h} &\leq \frac{1}{2}\ell \frac{1+\theta}{\theta} \\ \liminf_{h \rightarrow 0} \frac{\partial f(x+h) - Df(x)}{h} &\geq \frac{1}{2}\ell(1+\theta). \end{aligned}$$

These inequalities are valid for all  $\theta \in ]0, 1[$ , hence we have really proved that the  $\limsup$  and the  $\liminf$  are both equal to  $\ell$ .  $\square$

Note that the equivalence with the usual second derivative still does not hold: Example 4.1.9 is differentiable in the sense of Theorem 5.1.2 but not in the sense of (5.1.3), since  $Df(x + h)$  does not even exist in the neighborhood of 0. On the other hand, the property (5.1.1) appears as a suitable adaptation of (5.1.3) to the case of a “set-valued derivative”; therefore we agree to postulate Definition 5.1.1 as the *second differentiability* of a convex function. It is clear, for example from (4.1.7), that  $D_2f$  is a nonnegative number whenever it exists. Lebesgue’s differentiation theorem now says:

**Theorem 5.1.3** *A function  $f \in \text{Conv } \mathbb{R}$  is twice differentiable almost everywhere on the interior of its domain.*  $\square$

Unfortunately, this kind of second differentiability result does not help much in terms of  $f$ . Consider for example a piecewise affine function:

$$f(x) := \max \{s_j x - r_j : j = 1, \dots, m\}.$$

It has first and second derivatives except at a finite number of points (those where two different affine pieces meet, see Theorem 4.3.2). Its second derivative is 0 wherever it exists, but yet  $f$  differs substantially from being affine.

**Remark 5.1.4** The derivative  $Df$  of  $f \in \text{Conv } \mathbb{R}$  is locally integrable on the interior of its domain  $I$ ; as such, it can be seen as a distribution on  $I$ : why not consider its differentiation in the sense of distributions, then? A second derivative of  $f$  would be obtained, which would be a nonnegative Radon measure; for example, the second derivative of  $|\cdot|$  would be the Dirac measure at 0: the piecewise affine  $f$  above would be reconstructed with the sole help of this second derivative.

However, this approach is blind to sets of zero-measure; as such, it does not help much in optimization, where one is definitely interested in a designated point (the optimum): for this purpose, a *pointwise* differentiation is in order.  $\square$

## 5.2 One-Sided Second Derivatives

In Definition 5.1.1, existence of the usual first derivative is required at  $x$ , so as to control the difference quotient (5.1.1). However, we can get rid of this limitation; in fact, if  $h \downarrow 0$ , say, Theorem 4.2.1(iii) tells us that  $[\partial f(x + h) - D_+ f(x)]/h$  is the appropriate difference quotient – and the situation is symmetric for  $h \uparrow 0$ . The way is open to “half-second derivatives”. From now on, it is convenient to switch to the directional notation of Remark 4.1.4: for given  $x \in \text{int dom } f$ , we fix  $d \neq 0$  and we set  $h = td$ ,  $t > 0$ . We make appropriate substitutions in (5.1.1), (5.1.3) and (5.1.4) to obtain respectively

$$\lim_{t \downarrow 0} \frac{\partial f(x + td) - f'(x, d)}{t}, \quad (5.2.1)$$

$$\forall s(t) \in [D_- f(x + td), D_+ f(x + td)], \quad \lim_{t \downarrow 0} \frac{s(t) - f'(x, d)}{t}, \quad (5.2.2)$$

$$\lim_{t \downarrow 0} \frac{f(x + td) - f(x) - tf'(x, d)}{\frac{1}{2}t^2}. \quad (5.2.3)$$

As before, the definitions (5.2.1) and (5.2.2) are just equivalent: if one of the limits exists, the other two exist as well and are the same; this is the so-called point of view of Dini. As for (5.2.3) (the point of view of de la Vallée-Poussin), equivalence also holds:

**Theorem 5.2.1** *If one of the limits in (5.2.1)–(5.2.3) exists and is denoted by  $f''(x, d)$  ( $\geq 0$ ), then the other limits exist as well and are equal to  $f''(x, d)$ .*

PROOF. Just reproduce the proof of Theorem 5.1.2, without bothering with the sign of  $h$ .  $\square$

To illustrate what has been gained in passing from §5.1 to §5.2, take Example 4.1.9 and modify  $\varphi$  by setting  $\varphi(u) = 0$  for  $u \leq 0$ . Then the new  $f$  has the two “half-second derivatives”  $f''(0, -1) = 0$  and  $f''(0, 1) = 1$ .

Still, the limits in (5.2.1)–(5.2.2) may fail to exist, for two possible reasons: the difference quotients may go to  $+\infty$ , as in  $f(x + td) = t^{3/2}$ , or they may have several cluster points. Take again Example 4.1.9:  $\partial f(0+t)$  is squeezed between the curves  $s = t$  and  $s = t/(1+t)$ , which are tangent to each other at 0. If  $\varphi$  is modified so that this second curve becomes  $s = \frac{1}{2}t$ , say, then the set of cluster points in the difference quotient (5.2.1) blows up to the segment  $[1/2, 1]$ .

**Remark 5.2.2 (Interpretation of Second Difference Quotients)** For fixed  $x$  and  $d$ , consider the family of parabolas of equations indexed by  $c \geq 0$ :

$$\tau \mapsto p_c(\tau) = \frac{1}{2}c\tau^2 + s_0\tau + f(x), \quad \text{with } s_0 = f'(x, d). \quad (5.2.4)$$

They are constructed in such a way that  $p_c(0) = f(x)$  and  $p'_c(0) = f'(x, d)$ .

Now, fix  $t > 0$  and compute  $c$  so as to fit either the slope-value  $p'_c(t) = s(t)$  or the function-value  $p_c(t) = f(x + td)$ . In the first case,  $c$  is given by the difference quotient in (5.2.2) and in the second case by the difference quotient in (5.2.3). Both difference quotients thus appear as an estimate of the “curvature” of  $f$  at  $x$  in the direction  $d$ .  $\square$

### 5.3 How to Recognize a Convex Function

Given a function defined on an interval  $I$ , the question is now: can we decide whether it is convex on  $I$  or not? The answer depends on how much information is available: about the function itself, about its first derivatives (possibly one-sided), or about its second derivatives (or some sort of generalization). We review here the main criteria that are useful in optimization.

**(a) Using the Function Itself** Many criteria exist, relying on the definition of  $f$  and nothing more. Some of them are rather involved, most of them are of little relevance in the context of optimization. The most useful attitude is generally to view  $f$  as being *constructed* from other functions known to be convex, via operations such as those of §2.1 – and others to be seen in Chap. IV.

At this stage, the criterion 1.1.4 of increasing slopes should not be forgotten:  $f$  is convex if and only if the function

$$\Delta_1 f(x, x') := \frac{f(x) - f(x')}{x - x'}, \quad (5.3.1)$$

defined for pairs of different points in  $I$ , is increasing in each of its arguments. Note, however, that  $\Delta_1 f$  is a symmetric function of its two variables; hence it suffices that  $\Delta_1 f(x, \cdot)$  be increasing for each  $x$ .

As seen in §3.1, convexity of  $f$  on  $I = [a, b]$  implies its upper semi-continuity at  $a$  and  $b$ . Conversely, if  $f$  is convex on  $\text{int } I$ , and upper semi-continuous (relative to  $I$ ) on the boundary of  $I$ , then  $f$  is convex on  $I$ : just pass to the limit in (1.1.1). We will therefore content ourselves with checking the convexity of a given function on an *open* interval. Then, checking convexity on the closure of that interval will reduce to a study of continuity, usually much easier.

**(b) Using the First Derivative** Passing to the limit in  $\Delta_1 f$  of (5.3.1), one obtains the following result:

**Theorem 5.3.1** *Let  $f$  be continuous on an open interval  $I$  and possess an increasing right-derivative, or an increasing left-derivative, on  $I$ . Then  $f$  is convex on  $I$ .*

PROOF. Assume that  $f$  has an increasing right-derivative  $D_+ f$ . For  $x, x'$  in  $I$  with  $x < x'$  and  $u \in ]x, x'[$ , there holds

$$\frac{f(u) - f(x)}{u - x} \leq \sup_{t \in ]x, u[} D_+ f(t) \leq \inf_{t \in ]u, x'[} D_+ f(t) \leq \frac{f(x') - f(u)}{x' - u}$$

(the first and last inequalities come from mean-value theorems – in inequality form – for continuous functions admitting right-derivatives). Then (1.1.1) is obtained via a multiplication by  $x' - x > 0$ , knowing that  $u = \alpha x + (1 - \alpha)x'$  for some  $\alpha \in ]0, 1[$ . The proof for  $D_- f$  is just the same.  $\square$

**Corollary 5.3.2** *Assume that  $f$  is differentiable on  $I$  with an increasing derivative on an open interval  $I$ . Then  $f$  is convex on  $I$ .*  $\square$

**(c) Using the Second Derivative** To begin with, an immediate consequence of Corollary 5.3.2 is the following well-known criterion, by far the most useful of all, even though second differentiability is required:

**Theorem 5.3.3** *Assume that  $f$  is twice differentiable on an open interval  $I$ , and its second derivative is nonnegative on  $I$ . Then  $f$  is convex on  $I$ .*  $\square$

To illustrate a combination of Theorems 5.3.1 and 5.3.3, assume for example that  $f$  is “piecewise  $C^2$  with increasing slopes”, namely: there is a subdivision  $x_0 = a < x_1 < \dots < x_k = b$  of  $I = ]a, b[$  such that:

- $f$  is continuous on  $I$ ,
- $f$  is of class  $C^2$  and  $D_2 f \geq 0$  on each subinterval  $]x_{i-1}, x_i[$ ,  $i = 1, \dots, k$ ,
- $f$  has one-sided derivatives at  $x_1, \dots, x_{k-1}$  satisfying

$$D_- f(x_i) \leq D_+ f(x_i) \quad \text{for } i = 1, \dots, k-1.$$

Then  $f$  is convex on  $I$ .

In the absence of second differentiability, some sort of substitute is required to determine convexity. Translating to second order the criterion using the (symmetric) function  $\Delta_1 f$  of (5.3.1), we obtain:  $f$  is convex if and only if  $\Delta_2 f$  is nonnegative on  $I \times I \times I$ , where

$$\Delta_2 f(x, x', x'') := \frac{1}{x' - x''} \left[ \frac{f(x) - f(x')}{x - x'} - \frac{f(x) - f(x'')}{x - x''} \right]$$

is defined for all triples of different points  $x, x', x''$  in  $I$ . Note that  $\Delta_2 f$  is symmetric in its three variables.

Letting  $x'$  and  $x''$  tend to  $x$  in  $\Delta_2 f$ , just as was done with  $\Delta_1 f$ , one can get an analogue to Theorem 5.3.1. One must be careful when letting  $x'$  and  $x''$  converge, however: consider

$$x \mapsto f(x) := \min \left\{ \frac{1}{2}x^2 + x, \frac{1}{2}x^2 - x \right\}. \quad (5.3.2)$$

Its half-second derivatives (5.2.1) are constantly 1, but it is not convex: when passing to the limit with  $x'$  and  $x''$ , account must be taken of both sides of  $x$ . The “Schwarz second derivative”, for example, does the job by taking  $x - x' = x'' - x$ :

$$\bar{\Delta}_2 f(x) := \limsup_{t \downarrow 0} \frac{f(x-t) - 2f(x) + f(x+t)}{t^2}. \quad (5.3.3)$$

We obtain the second derivative of  $f$  at  $x$  if there is one; the counter-example (5.3.2) has  $\bar{\Delta}_2 f(0) = -\infty$ , and is thus eliminated. When  $f$  is convex,

$$\bar{\Delta}_2 f(x) \geq 0 \quad \text{for all } x \in I. \quad (5.3.4)$$

This condition turns out to be sufficient if combined with the continuity of  $f$ :

**Theorem 5.3.4** *Assume that  $f$  is continuous on the open interval  $I$  and that (5.3.4) holds. Then  $f$  is convex on  $I$ .*

PROOF. Take  $a$  and  $b$  in  $I$  with  $a < b$ ,  $\alpha \in ]0, 1[$  and set  $x := \alpha a + (1 - \alpha)b$ . We have to prove the “mean-value inequality”

$$f(x) \leq f(a) + \frac{f(b) - f(a)}{b - a}(x - a). \quad (5.3.5)$$

We take

$$g(x) := f(x) - f(a) - \frac{f(b) - f(a)}{b - a}(x - a),$$

and we prove  $g \leq 0$  on  $]a, b[$ . We have  $g(a) = g(b) = 0$  and, since  $f$  and  $g$  differ by an affine function,  $\bar{\Delta}_2 g = \bar{\Delta}_2 f$ .

Suppose first

$$\bar{\Delta}_2 g(x) = \bar{\Delta}_2 f(x) > 0 \quad \text{for all } x \in ]a, b[. \quad (5.3.6)$$

We claim that  $g$  is then nonpositive on  $]a, b[$ : if such were not the case, the continuous  $g$  would assume its maximal value at some  $x^* \in ]a, b[$  and the relation

$$g(x^* - t) - 2g(x^*) + g(x^* + t) < 0 \quad \text{for all } t \text{ small enough}$$

would contradict (5.3.6). Thus (5.3.5) is proved.

Now define  $f_k(x) := f(x) + 1/k x^2$ . If (5.3.4) holds,  $\bar{\Delta}_2 f_k$  is positive on  $]a, b[$  and, from the first part of the proof,  $f_k$  is convex. Its pointwise limit  $f$  is therefore convex (Proposition 2.2.1).  $\square$

**Remark 5.3.5** With relation to Remark 5.2.2, observe that the difference quotient in (5.3.3) represents one more “curvature” estimate. Let  $s_0$  be free in (5.2.4) and force  $p_c$  to coincide with  $f$  at  $x, x - t, x + t$ : we again obtain  $c = \Delta_2 f(x, x - t, x + t)$ .  $\square$

## 6 First Steps into the Theory of Conjugate Functions

On several occasions, we have encountered the *conjugate function* of  $f$ , defined by

$$\mathbb{R} \ni s \mapsto f^*(s) := \sup \{sx - f(x) : x \in \text{dom } f\}. \quad (6.0.1)$$

Because  $sx$  is a finite number, we can let  $x$  run through the whole of  $\mathbb{R}$ , and of course this does not change the supremum: instead of (6.0.1), we may as well write the simpler form

$$f^*(s) = \sup_{x \in \mathbb{R}} [sx - f(x)]. \quad (6.0.2)$$

**Remark 6.0.1** Some comments are in order with respect to Remark 1.3.3. What is actually computed in (6.0.2) is

$$\inf_x [f(x) - sx], \quad (6.0.3)$$

a number which is certainly not  $+\infty$ . As a result, its *opposite*  $f^*(s)$  is in our space of interest  $\mathbb{R} \cup \{+\infty\}$ . Furthermore, this opposite will be seen to behave as a convex function of  $s$  (already here, remember Proposition 2.1.2).

Indeed, one should realize that (6.0.1), (6.0.2) – or (6.0.3) – actually means

$$f^*(s) = \sup \{sx - r : (x, r) \in \text{epi } f\}, \quad (6.0.4)$$

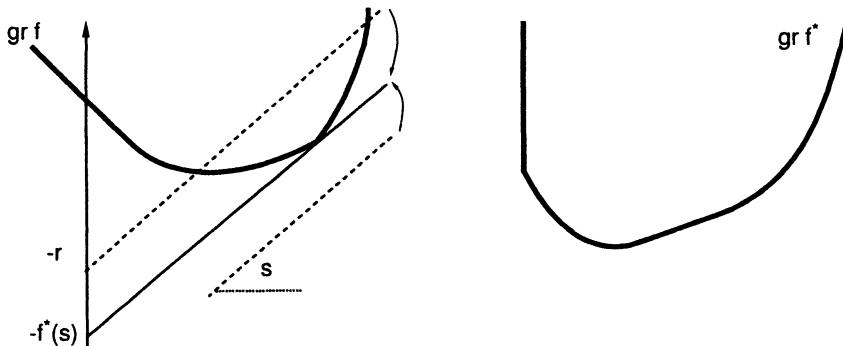


Fig. 6.0.1. Constructing a conjugate function

and this last writing has two advantages: first, it suppresses the “ $-f(x)$ ” operation; and more importantly, it interprets the conjugacy operation as the *supremum of a linear function*  $[\ell_s(x, r) := sx - r]$  over a closed convex set of  $\mathbb{R}^2$ . We will return later (Chapters IV and X) to this aspect; considering that (6.0.4) is rather heavy, the versions (6.0.1) or (6.0.2) are generally preferable, and will be generally preferred.

□

We retain from (6.0.4) the geometrical interpretation displayed in Fig. 6.0.1: for given  $s$  and  $r$ , consider the affine function  $a_{s,r}$  defined by

$$\mathbb{R} \ni x \mapsto a_{s,r}(x) = sx - r$$

and the corresponding line  $\text{gr } a_{s,r}$  in  $\mathbb{R}^2$ . Due to the geometry of an epigraph, there are two kinds of  $r$  for given  $s$ : those, small enough, such that  $a_{s,r} \leqslant f$ ; and those so large that  $a_{s,r}(x) > f(x)$  for some  $x$ . The particular  $r = f^*(s)$  is their common bound, obtained when the line  $\text{gr } a_{s,r}$  “leans” on  $\text{epi } f$ , or *supports*  $\text{epi } f$ . For the particular value  $s = 0$ , we obtain

$$-f^*(0) = \inf \{f(x) : x \in \mathbb{R}\}. \quad (6.0.5)$$

Figure 6.0.1 displays the set for which  $f^*$  is finite; and this set depends exclusively on the behaviour of  $f$  at infinity, which therefore plays an important role for the determination of  $\text{dom } f^*$  (remember §2.3). On the other hand, let  $x_0 \in \text{dom } f$  and choose  $s \in \partial f(x_0)$ ; then the corresponding “optimal” line supports  $\text{gr } f$  at  $(x_0, f(x_0))$ , so that  $f^*(s) = sx_0 - f(x_0)$  for such an  $s$ .

**Examples 6.0.2** For each  $f \in \text{Conv } \mathbb{R}$  considered below, we give the corresponding conjugate function  $f^*$ . Draw the graph of  $f^*$  in each case.

- $f(x) = |x|$ : then  $f^*$  is the indicator function  $I_{[-1, +1]}$ ; more simply,  $f(x) = sx$  gives  $f^* = I_{\{s\}}$ .
- $f(x) = (1/p)|x|^p$ , with  $p > 1$ : then  $f^*(s) = (1/q)|s|^q$ , with  $1/p + 1/q = 1$ . In particular,  $f^* = f$  if  $p = 2$ .
- $f(x) = x \log x$  if  $x > 0$ ,  $+\infty$  if not: then  $f^*(s) = \exp s - 1$  for all  $s \in \mathbb{R}$ .

–  $f(x) = -\sqrt{1-x^2}$  if  $|x| \leq 1$ ,  $+\infty$  if not (the ball-pen function of Fig. 2.1.1): then  $f^*(s) = \frac{s}{\sqrt{1+s^2}}$ .  $\square$

It is important to realize that the argument  $s$ , which  $f^*$  depends on, is a *slope*, i.e. strictly speaking an element of the *dual* of  $\mathbb{R}$ . When taking again the conjugate of  $f^*$ , one goes back to the primal and the result is the *biconjugate* function of  $f$ :

$$f^{**}(x) := (f^*)^*(x) = \sup \{sx - f^*(s) : s \in \text{dom } f^*\}.$$

For illustration, compute the biconjugates in the examples above.

The transformation  $f \mapsto f^*$  is (the one-dimensional version of) the so-called *Fenchel correspondence*, and is closely related to the *Legendre transform*. In view of its importance for a deep understanding of the properties of a convex function, we are going to explore step by step some basic results about it.

## 6.1 Basic Properties of the Conjugate

First of all, the very definition (6.0.1) directly implies the relation

$$sx \leq f(x) + f^*(s) \quad \text{for all } x \in \text{dom } f \text{ and all } s \in \text{dom } f^*, \quad (6.1.1)$$

called the *Young-Fenchel inequality* (which, incidentally, holds for all  $s$  and  $x$ !).

**Proposition 6.1.1** *Let  $f \in \text{Conv } \mathbb{R}$ . Then*

- *the conjugate of  $f$  is a closed convex function ( $f^* \in \text{Conv } \mathbb{R}$ ),*
- *the biconjugate of  $f$  is its closure ( $f^{**} = \text{cl } f$ ).*

PROOF. The function  $f^*$  takes its values in  $\mathbb{R} \cup \{+\infty\}$  by construction. Its domain is nonempty, see Remark 4.1.7. Then, its convexity and closedness result from Proposition 3.3.2.

Now, use the form (6.0.4) to define  $f^{**}$ :

$$f^{**}(x) = \sup_{s,r} \{sx - r : r \geq f^*(s)\}. \quad (6.1.2)$$

By definition of  $f^*$ , to say  $r \geq f^*(s)$  is to say that, for all  $y \in \text{dom } f$ ,

$$r \geq sy - f(y), \quad \text{i.e.} \quad sy - r \leq f(y).$$

In other words, (6.1.2) can be written

$$f^{**}(x) = \sup_{s,r} \{sx - r : sy - r \leq f(y) \quad \text{for all } y \in \text{dom } f\},$$

in which we recognize the expression (3.2.4) of  $\text{cl } f$ .  $\square$

When conjugating a function  $f$ , one considers the set of all affine functions minorizing it. As mentioned in Remark 3.2.6, this is also the set of all affine functions minorizing  $\text{cl } f$ . It follows that  $f$  and  $\text{cl } f$  have the same conjugate: from now on, we may assume that the convex  $f$  is closed, this will be good enough. Then the relation  $f^{**} = f$ , established in Proposition 6.1.1, shows that the Legendre-Fenchel transformation is an *involution* in  $\text{Conv } \mathbb{R}$ . This is confirmed by the next result, in which we have also an involution between  $s$  and  $x$  via the solution-set of (6.0.1).

**Proposition 6.1.2** *Let  $f \in \text{Conv } \mathbb{R}$ . Then*

$$sx = f(x) + f^*(s) \quad \text{if and only if} \quad x \in \text{dom } f \text{ and } s \in \partial f(x); \quad (6.1.3)$$

$$s \in \partial f(x) \quad \text{if and only if} \quad x \in \partial f^*(s). \quad (6.1.4)$$

PROOF. We have that

$$-f^*(s) = \inf \{f(x) - sx : x \in \mathbb{R}\}.$$

The function  $g_s : x \mapsto f(x) - sx$ , which is in  $\text{Conv } \mathbb{R}$ , achieves its infimum at  $\bar{x}$  if and only if  $0 \in \partial g(\bar{x})$  – see (4.1.8) – i.e.

$$-f^*(s) = f(\bar{x}) - s\bar{x} \quad \text{if and only if} \quad s \in \partial f(\bar{x}).$$

This implies  $s \in \text{dom } f^*$  and can be written as (6.1.3). Applying this same result to  $f^*$  (which is closed), we obtain

$$x \in \partial f^*(s) \quad \text{if and only if} \quad sx = f^*(s) + f^{**}(x),$$

which is again (6.1.3) since  $f^{**} = f$ .  $\square$

What (6.1.3) says is that the pairs  $(x, s) \in \mathbb{R}^2$  for which the inequality of Young-Fenchel (6.1.1) holds as an equality form exactly the graph of  $\partial f$ . In view of (6.1.4), the mapping  $\partial f^*$  is obtained by *inverting* the mapping  $\partial f$ , i.e. reflecting its graph across the line of equation  $s = x$ : see Fig. 6.1.1, and remember the increasing property (4.2.1).

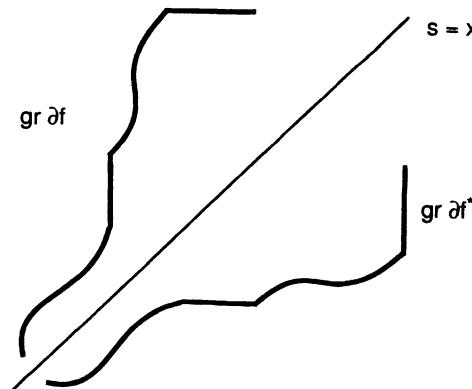


Fig. 6.1.1. The symmetry between  $\partial f$  and  $\partial f^*$

**Remark 6.1.3** The above inversion property suggests a way of computing a conjugate which may be useful: “differentiate”  $f$  to obtain  $\partial f$ ; then invert the result and integrate it to obtain  $f^*$  up to a constant. As an exercise, compute graphically the conjugate of  $\frac{1}{2}x^2 + |x|$ .  $\square$

## 6.2 Differentiation of the Conjugate

The question we address in this section is: what differentiability properties can be expected for  $f^*$ , perhaps requiring from  $f$  something more than mere convexity?

Let  $s_0 \in \text{int dom } f^*$  and consider the statement

$$f^* \text{ is differentiable at } s_0.$$

According to Proposition 6.1.2, it means

$$\text{there is a unique solution to the "equation" } (\text{in } x) \partial f(x) \ni s_0, \quad (6.2.1)$$

which in turn relies on the key property

$$\partial f \text{ is "strictly increasing" on its domain,} \quad (6.2.2)$$

in the sense that  $\partial f(x_1) < \partial f(x_2)$  whenever  $x_1 < x_2$ . As is easily checked, this last property is equivalent to

$$f \text{ is strictly convex.} \quad (6.2.3)$$

Thus, we have:

**Proposition 6.2.1** *Let  $f$  be strictly convex. Then  $f^*$  is differentiable on the interior of its domain and, for all  $s \in \text{int dom } f^*$ ,*

$$Df^*(s) = x(s)$$

where  $x(s)$  is the unique solution of

$$s \in \partial f(x), \quad \text{or} \quad sx - f(x) = f^*(s), \quad \text{or} \quad \min_x [f(x) - sx]. \quad \square$$

The converse to Proposition 6.2.1 is false:  $f^*$  may be differentiable on the interior of its domain while  $f$  is not strictly convex. A counter-example is

$$f_{(1)}(x) := \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq 1, \\ |x| - 1/2 & \text{if } |x| \geq 1, \end{cases} \quad (6.2.4)$$

for which easy computations give

$$f_{(1)}^*(s) = \begin{cases} \frac{1}{2}s^2 & \text{if } |s| \leq 1, \\ \frac{1}{2}s^2 + I_{[-1, +1]}(s) = +\infty & \text{if } |s| > 1. \end{cases}$$

The only explanation is that (6.2.1) (assumed to hold for all  $s_0 \in \text{int dom } f^*$ ) does not imply (6.2.2). More precisely, two different  $x_1$  and  $x_2$  are allowed to give a nonempty intersection  $\partial f(x_1) \cap \partial f(x_2) \ni s_0$ , provided that this  $s_0$  falls on the boundary of  $\text{dom } f^*$ . Some additional assumption is necessary to rule this case out; among other things, the following result illustrates further the involutonal character of the Legendre-Fenchel transformation.

**Proposition 6.2.2** *Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be strictly convex, differentiable, and 1-coercive ( $f(x)/|x| \rightarrow +\infty$  for  $|x| \rightarrow \infty$ ). Then*

(i)  $f^*$  enjoys the same properties

and, for all  $s \in \mathbb{R}$ ,

(ii) there is a unique solution to the equation  $Df(x) = s$ ,

(iii)  $f^*(s) = s(Df)^{-1}(s) - f((Df)^{-1}s)$ .

PROOF. We claim first that the 1-coercivity assumption on  $f$  (which, according to (2.3.3), is equivalent to  $f'_\infty(1) = f'_\infty(-1) = +\infty$ ) amounts to saying that

$$\lim_{x \rightarrow +\infty} Df(x) = - \lim_{x \rightarrow -\infty} Df(x) = +\infty.$$

In fact, for  $x > 0$ , (4.1.7) gives

$$Df(x) \geq \frac{f(x) - f(0)}{x}.$$

When  $x \rightarrow +\infty$ , the right-hand side goes to  $f'_\infty(1)$ , so  $f'_\infty(1) = +\infty$  implies  $Df(x) \rightarrow +\infty$ . To prove the converse, let  $x \rightarrow +\infty$  in the inequalities

$$Df(x) \leq f(x+1) - f(x) \leq f'_\infty(1),$$

which come from the property of increasing slopes. The same proof works for  $x \rightarrow -\infty$  and establishes our claim.

Remembering the equivalence between (6.2.2) and (6.2.3), we therefore see that  $Df$  is a bijection from  $\mathbb{R}$  onto  $\mathbb{R}$ . Its inverse  $(Df)^{-1} = Df^*$  is a bijection as well and the whole result follows.  $\square$

**Example 6.2.3** The function  $f(x) = \operatorname{ch} x$  satisfies the assumptions of Proposition 6.2.2:  $Df(x) = \operatorname{sh} x$ , hence the inverse  $Df^*(s) = (\operatorname{sh})^{-1}(s)$ . We readily obtain

$$f^*(s) = s(\operatorname{sh})^{-1}(s) - \sqrt{1+s^2},$$

which is an illustration of (iii). Among other things, the 1-coercivity of the above function is implied by (i), but could not be seen at first glance.  $\square$

Consider now the problem of differentiating  $f^*$  twice, which is (not unexpectedly) more complex. To get an idea of what can be expected and what is hopeless, we suggest meditating on the following examples.

#### Examples 6.2.4

(a)  $f_1 = |\cdot|$  is  $C^\infty$  in a neighborhood of an arbitrary  $x_0 > 0$ . Nevertheless,  $f_1^* = I_{[-1,+1]}$  is not even finite in a neighborhood of  $s_0 = Df_1(x_0)$  ( $= 1$  for all  $x_0 > 0$ ).

(b) The previous function was not differentiable everywhere, but consider

$$f_2(x) = \begin{cases} 0 & \text{if } |x| \leq 1, \\ \frac{1}{2}(|x|-1)^2 & \text{otherwise.} \end{cases}$$

Then,  $f_2^*(s) = 1/2s^2 + |s|$  is still not differentiable (at  $s = 0$ ).

(c) The following function is convex, 1-coercive and twice differentiable everywhere:

$$f_3(x) = \begin{cases} 0 & \text{if } |x| \leq 1, \\ \frac{1}{3}(|x| - 1)^3 & \text{otherwise.} \end{cases}$$

Yet,  $f_3^*(s) = 2/3 |s|^{3/2} + |s|$  is not even once differentiable (at  $s = 0$ ).

(d) A slight perturbation of the previous example is  $f_4(x) = 1/3 |x - 1|^3$ ; it is strictly convex, 1-coercive and twice differentiable throughout  $\mathbb{R}$  but  $f_4^*(s) = 2/3 |s|^{3/2} + s$  is only *once* differentiable.

(e) Take the conjugate of (6.2.4):  $f_5(x) = 1/2 x^2 + I_{[-1, +1]}(x)$  is  $C^\infty$  on the interior of its domain, with  $D_2 f_5 > 0$  throughout (while  $D_2 f_4$  was 0 at the only point 0). Its conjugate  $f_5^* = f_{(1)}$  of (6.2.4) is not twice differentiable at  $\pm 1$ .  $\square$

The deep reason for all these oddities is that  $f^*$  is a *global* concept, as it takes into account a priori the behaviour of  $f$  on its whole domain; as a result, the smoothness of  $f^*$  is a tricky matter. We just mention two results: a *local* one, and a *global* one which echoes Proposition 6.2.1 via the inverse function theorem.

**Proposition 6.2.5** *Assume that  $f \in \text{Conv } \mathbb{R}$  is twice differentiable at  $x_0$  (in the sense of Definition 5.1.1) with  $D_2 f(x_0) > 0$ . Then  $f^*$  is likewise twice differentiable at  $s_0 = Df(x_0)$  and*

$$D_2 f^*(s_0) = \frac{1}{D_2 f(x_0)}.$$

PROOF. First of all, we claim that  $f^*$  is differentiable at  $s_0$ , with derivative  $x_0$ . In fact,  $x_0 \in \partial f^*(s_0)$  because of (6.1.4). If the convex set  $\partial f^*(s_0)$  contains another  $x_0 + d$ , then it contains also the whole interval  $x_0 + [0, 1]d$ : we have  $s_0 \in \partial f(x_0 + td)$  for  $t \downarrow 0$ ; comparing with (5.1.1), we see that the positivity of  $D_2 f(x_0)$  is contradicted.

Now, we want to prove that, for arbitrary  $x \in \partial f^*(s)$ ,

$$\frac{x - x_0}{s - s_0} \rightarrow \frac{1}{D_2 f(x_0)} > 0, \quad \text{i.e.} \quad \frac{s - s_0}{x - x_0} \rightarrow D_2 f(x_0)$$

when  $s \rightarrow s_0$ ; but this follows from (5.1.1):  $s \in \partial f(x)$  and Corollary 4.2.3 tells us that  $x \rightarrow x_0$ , since  $f^*$  is differentiable at  $x_0$ .  $\square$

As a result, suppose that  $f \in \text{Conv } \mathbb{R}$  is twice differentiable on  $\text{int dom } f$  with  $D_2 f > 0$  throughout. Then  $f^*$  enjoys the same properties, but *only* on the image-set  $Df(\text{int dom } f)$ ; see Example 6.2.4(e).

A one-sided version of Proposition 6.2.5 can also be stated just as in Theorem 5.2.1. We rather give the global version below, obtained via the  $C^1$  parametrization of Proposition 6.2.1:  $Df^* = (Df)^{-1}$ .

**Corollary 6.2.6** *Assume that  $f$  is 1-coercive, and twice differentiable on  $\mathbb{R}$ , with  $D_2 f > 0$  throughout. Then  $f^*$  is likewise and*

$$D_2 f^* = \frac{1}{D_2 f \circ (Df)^{-1}}.$$

For illustration, see again Example 6.2.3:

$$D_2 f(x) = \text{ch } x \quad \text{and} \quad D_2 f^*(s) = \sqrt{1 + s^2}.$$

### 6.3 Calculus Rules with Conjugacy

In §2.1, we have introduced some operations preserving convexity, whose effect on the subdifferentials has been seen in §4.3. Here, we briefly review their effect on the conjugate function.

**Proposition 6.3.1** *Let  $f_1$  and  $f_2$  be two (closed) convex functions, minorized by a common affine function. Then*

$$(f_1 \downarrow f_2)^* = f_1^* + f_2^*. \quad (6.3.1)$$

PROOF. The proof illustrates some properties of extremization (see in particular §A.1.2). For  $s \in \mathbb{R}$ ,

$$\begin{aligned} (f_1 \downarrow f_2)^*(s) &= \sup_x \{sx - \inf_{x_1+x_2=x} [f_1(x_1) + f_2(x_2)]\} \\ &= \sup_{x_1+x_2=x} [s(x_1 + x_2) - f_1(x_1) - f_2(x_2)] \\ &= \sup_{x_1, x_2} [s(x_1 + x_2) - f_1(x_1) - f_2(x_2)] \\ &= \sup_{x_1} [sx_1 - f_1(x_1)] + \sup_{x_2} [sx_2 - f_2(x_2)] \end{aligned}$$

and we recognize  $f_1^*(s) + f_2^*(s)$  in this last expression.  $\square$

The dual version of this result is that, if  $f_1$  and  $f_2$  are two closed convex functions finite at some common point, then

$$(f_1 + f_2)^* = f_1^* \downarrow f_2^*. \quad (6.3.2)$$

The way to prove it is to observe that the two functions  $f_1^*$  and  $f_2^*$  satisfy the assumptions of Proposition 6.3.1, and their conjugates are  $f_1$  and  $f_2$  respectively; hence

$$(f_1^* \downarrow f_2^*)^* = f_1 + f_2.$$

Taking the conjugate of both sides and knowing that an infimal convolution is closed (see Remark 3.3.4) gives directly (6.3.2). In several dimensions, however, an inf-convolution is no longer closed, so technical difficulties can be anticipated to establish (6.3.2).

The value at  $s = 0$  of the function (6.3.2) gives an interesting relation: in view of (6.0.5), we have

$$\inf_{x \in \mathbb{R}} [f_1(x) + f_2(x)] = -(f_1 + f_2)^*(0) = \inf_{s \in \mathbb{R}} [f_1^*(s) + f_2^*(-s)],$$

which is known as (the univariate version of) *Fenchel's duality theorem* – but once again, beware that it does not extend readily to several variables.

Formulae (6.3.1) and (6.3.2) show that the addition of functions and their infimal convolution are operations dual to each other. The sup-operation is more complex: it is dual to an operation that we have not seen yet, namely that of taking the *closed convex hull* of a nonconvex function. Indeed, convexity of  $f$  is by no means necessary to define its conjugate (6.0.1): the result is “meaningful” as soon as we have:

- (i)  $f$  is not identically  $+\infty$  (otherwise  $f^*$  would be – identically! –  $-\infty$ )
- (ii)  $f$  is minorized by some affine function (otherwise  $f^*$  would be identically  $+\infty$ ).

Now, to  $f$  satisfying these properties, we can associate the family of affine functions  $s \mapsto sx - f(x)$ , indexed by  $x \in \mathbb{R}$ : Proposition 3.3.2 tells us that their supremum  $f^*$  is a closed convex function of  $s$ .

In a word, the conjugacy operation can perfectly well be applied to any function  $f$  satisfying the conditions (i) and (ii) above, “and nothing more”. Looking again at the proof of Proposition 6.1.1, we see that the biconjugate of  $f$  is then the pointwise supremum of all the affine functions minorizing  $f$ . The epigraph of  $f^{**}$  appears as the closed convex hull of  $\text{epi } f$ , as indicated by Fig. 6.3.1. In view of this remark, a more suggestive notation can be used:

$$f^{**} = \text{cl co } f = \overline{\text{co}} f. \quad (6.3.3)$$

This last function appears as the “close-convexification” of  $f$ , i.e. the largest closed and convex function minorizing  $f$ ; naturally,  $\overline{\text{co}} f \leq f$ !

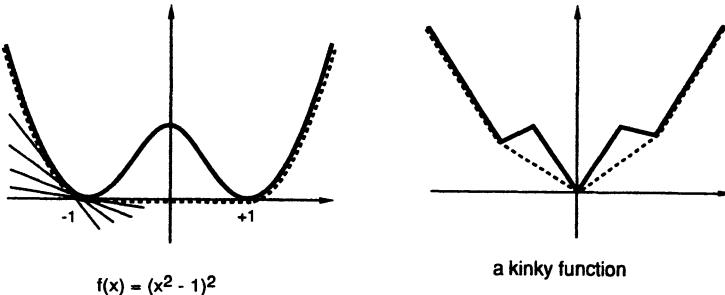


Fig. 6.3.1. Taking a closed convex hull

The extension thus introduced for the conjugacy is used in our next results.

**Proposition 6.3.2** *Let  $\{f_j\}_{j \in J}$  be a collection of functions not identically  $+\infty$ , and all minorized by some common affine function. Then the function  $f := \inf_{j \in J} f_j$  satisfies (i) and (ii), and its conjugate is*

$$(\inf_j f_j)^* = \sup_j (f_j^*). \quad (6.3.4)$$

PROOF. That  $f$  satisfies (i) and (ii) is clear. Then (6.3.4) is proved as (6.3.1), via the same properties of extremization.  $\square$

**Corollary 6.3.3** *Let  $\{g_j\}_{j \in J}$  be a collection of functions in  $\overline{\text{Conv}} \mathbb{R}$ , and suppose that there is some  $x_0$  such that  $\sup_{j \in J} g_j(x_0) < +\infty$ . Then*

$$(\sup_j g_j)^* = \overline{\text{co}}(\inf_j g_j^*).$$

PROOF. Proposition 6.3.2 applied to  $f_j = g_j^*$  gives

$$(\inf_j g_j^*)^* = \sup_j g_j^{**} = \sup_j g_j.$$

The result follows from (6.3.3), by taking the conjugate of each side.  $\square$

**Example 6.3.4** Given two arbitrary functions  $\varphi$  and  $c$  from some arbitrary set  $Y$  to  $\mathbb{R}$ , consider the (closed and convex) function

$$\mathbb{R} \ni x \mapsto g(x) := \sup \{x c(y) - \varphi(y) : y \in Y\} \quad (6.3.5)$$

which we assume  $< +\infty$  for some  $x_0 \in \mathbb{R}$ . With the help of the notation

$$g_y(x) := x c(y) - \varphi(y) \quad \text{for all } y \in Y \text{ and } x \in \mathbb{R},$$

we can apply Proposition 6.3.3 to compute  $g^*$ . We directly obtain

$$g^*(s) = \overline{\text{co}}[\inf_{y \in Y} g_y^*(s)], \quad (6.3.6)$$

where the conjugate of each  $g_y$  is easy to compute:

$$g_y^*(s) = \begin{cases} \varphi(y) & \text{if } s = c(y) \\ \sup_x [(s - c(y))x + \varphi(y)] = +\infty & \text{otherwise.} \end{cases}$$

This calculation is of interest in optimization: consider the (abstract) minimization problem with one constraint

$$\left| \begin{array}{ll} \inf \varphi(y) & y \in Y \\ c(y) = s. & \end{array} \right. \quad (6.3.7)$$

Here, the right-hand side of the constraint is parametrized by  $s \in \mathbb{R}$ . The optimal value is a function of the parameter, say  $P(s)$ , usually called the *value-function*, or also primal, perturbation, or marginal function. Clearly enough, this function can be written

$$P(s) = \inf \{g_y^*(s) : y \in Y\}.$$

Observe that  $P$  has no special structure since we have made no assumption on  $Y, \varphi, c$  – other than  $g \not\equiv +\infty$  in (6.3.5). Nevertheless, what (6.3.6) tells us is that the closed convex hull of  $P$  is the conjugate of  $g$  in (6.3.5):

$$g^* = \overline{\text{co}} P.$$

In particular, if  $P$  happens to be closed and convex, we obtain from (6.0.5):  $-\inf g = g^*(0) = P(0)$ . With notation closer to that of (6.3.7), this means

$$\sup_{x \in \mathbb{R}} \inf_{y \in Y} [\varphi(y) - xc(y)] = \inf \{\varphi(y) : c(y) = 0\}. \quad \square$$

The closed convex hull of a function is an important object for optimization, even though it is not easily computable. A reason is that minimizing  $f$  or minimizing  $\overline{\text{co}} f$  are “equivalent” problems in the sense that:

$$\bar{x} \text{ minimizes } f \iff [\bar{x} \text{ minimizes } \overline{\text{co}} f \text{ and } \overline{\text{co}} f(\bar{x}) = f(\bar{x})].$$

Even more can be said:

**Theorem 6.3.5** *Let  $f : \mathbb{R} \rightarrow \mathbb{R}$  be a differentiable function with derivative  $Df$ . Then  $\bar{x}$  minimizes  $f$  on  $\mathbb{R}$  if and only if*

$$Df(\bar{x}) = 0 \quad \text{and} \quad \overline{\text{co}} f(\bar{x}) = f(\bar{x}).$$

*In such a case,  $\overline{\text{co}} f$  is differentiable and minimal at  $\bar{x}$ .*

PROOF. The condition  $Df(\bar{x}) = 0$  is known to be necessary for  $\bar{x}$  to minimize the differentiable function  $f$ . Furthermore, the (constant) affine function defined by  $\ell(x) \equiv f(\bar{x})$  minorizes  $f$  – hence  $\ell \leq \text{co } f$  – and coincides with  $f$  at  $\bar{x}$  – hence  $\ell(\bar{x}) = \text{co } f(\bar{x})$ .

Conversely, let  $x$  satisfy  $Df(x) = 0$  and  $\text{co } f(x) = f(x)$ . Since  $\text{co } f \leq f$ , we have

$$\frac{\text{co } f(x+h) - \text{co } f(x)}{h} \leq \frac{f(x+h) - f(x)}{h} \quad \text{for all } h > 0.$$

Letting  $h \downarrow 0$ , we obtain

$$D_+ \text{co } f(x) \leq Df(x) = 0.$$

Taking  $h < 0$ , we show likewise that

$$D_- \text{co } f(x) \geq Df(x) = 0.$$

On the other hand, the convex  $\text{co } f$  satisfies  $D_- \text{co } f \leq D_+ \text{co } f$ : we conclude that  $D \text{co } f(x) = 0$ ,  $\text{co } f$  has a 0-derivative at  $x$ , is therefore minimal at  $x$ , and  $f$  as well.  $\square$

Thus, what is needed for a stationary point  $x$  of  $f$  to be a minimum is just to satisfy  $\text{co } f(x) = f(x)$ . The examples of Fig. 6.3.1 help understanding this last property: the function  $(x^2 - 1)^2$  has the minima  $\pm 1$ , and 0 is left out. It is interesting to note that the condition  $Df(x) = 0$  is purely *local* and makes no reference whatsoever to minimality of  $x$ , rather than maximality, say. In fact, suppose  $f$  has only one-sided derivatives; if the stationarity condition “ $Df(x) = 0$ ” is replaced by the apparently natural “ $D_- f(x) \leq 0 \leq D_+ f(x)$ ”, then Theorem 6.3.5 breaks down: see the right part of Fig. 6.3.1. By contrast, the condition “ $\text{co } f(x) = f(x)$ ” has *global* character.

## II. Introduction to Optimization Algorithms

**Prerequisites.** Some knowledge of computer programming; elementary differential calculus in  $\mathbb{R}^n$ : inner products, gradient vectors, Hessian operators.

**Introduction.** In this chapter, we survey the techniques that are suitable for solving minimization problems. By “solving”, we mean actually computing a solution, or at least approximating it. In this domain, convexity of the functions involved is of little relevance, what is important is rather their *smoothness*. We will therefore limit our attention to smooth enough functions (say  $C^\infty$ ) and neglect their convexity as a minor detail: it will become important only in subsequent chapters.

Our aim is not to give a complete list of existing algorithms (there are other books for that); rather, we will extract those concepts that will be useful for us later, when we develop algorithms in the context of convex analysis.

### 1 Generalities

#### 1.1 The Problem

We are interested in the following optimization problem: given an *objective function*  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , find  $\bar{x} \in \mathbb{R}^n$  such that

$$\bar{f} := f(\bar{x}) \leq f(x) \quad \text{for all } x \in \mathbb{R}^n, \quad (1.1.1)$$

i.e. we want to solve (see the appendix §A.1.3 – A.1.4 for the notation)

$$\min \{f(x) : x \in \mathbb{R}^n\} =: \bar{f}. \quad (1.1.2)$$

**Remark 1.1.1** There are problems in which one is more interested in finding  $\bar{f}$ , and the precise value of  $\bar{x}$  is of smaller interest. This is the case for instance when  $f$  represents an actual cost, say in French francs, but (1.1.2) represents only a simulation (say of a power plant, an investment, ...) rather than the actual operation (of that power plant, ...). In some other cases, one is definitely more interested in finding  $\bar{x}$ , the value  $\bar{f}$  being only a by-product. This latter case is illustrated by the following well-known situation: to solve a system of equations

$$F(x) = 0 \quad \text{with} \quad F : \mathbb{R}^n \rightarrow \mathbb{R}^m, \quad (1.1.3)$$

one sometimes prefers to solve the optimization problem

$$\min \{\|F(x)\| : x \in \mathbb{R}^n\}$$

which is normally equivalent to (1.1.3). Clearly then, one is not very interested in the minimal *value*, which is normally 0, but more so in the minimal *solution*, which supposedly solves the original problem (1.1.3).  $\square$

- (i) *Existence* questions for (1.1.1) are rather well known. Provided that  $r \in \mathbb{R}$  is large enough to imply that the sublevel-set  $S_r(f)$  is nonempty – say  $r \geq f(x_1)$ ,  $x_1$  arbitrary – (1.1.2) is clearly equivalent to

$$\min \{f(x) : x \in S_r(f)\}.$$

If we assume that some  $S_r(f)$  is bounded, then the continuity of  $f$ , or even its lower semi-continuity, implies the existence of a solution  $\bar{x}$  to (1.1.2).

- (ii) *Uniqueness* is a different matter, for which the suitable assumptions are related with convexity: for example, strict convexity of  $f$  implies uniqueness of  $\bar{x}$  solving (1.1.2).
- (iii) *Recognizing* a solution is, after existence and uniqueness, the next question concerning (1.1.2): when can it be ascertained that a given  $\bar{x}$  is optimal? This is an unsolved problem, unless  $f$  has some particular structure – and convexity is again the most classical one (we neglect the direct checking of (1.1.1) for all possible  $x$ !). This means that the question of whether a given  $\bar{x}$  minimizes  $f$  cannot in general be answered with certainty. This question is important in practice, however, because optimization methods that actually compute an optimal solution will precisely be based on the characterization of such a solution. Thus, instead of really solving (1.1.2), one looks rather for a so-called local minimum, i.e. a point such that one must, to obtain better values for  $f$ , move a definite distance from it:

**Definition 1.1.2** A *local minimum* of  $f$  is an  $\bar{x} \in \mathbb{R}^n$  satisfying

$$\exists \varepsilon > 0 \quad \text{such that} \quad \|x - \bar{x}\| \leq \varepsilon \implies f(\bar{x}) \leq f(x). \quad (1.1.4)$$

$\square$

Now, differential properties of  $f$  help describing a local minimum, and the following result is classical:

**Theorem 1.1.3** Suppose  $f$  is a differentiable function.

- (a) First-order necessary condition: if  $\bar{x}$  is a local minimum then

$$\nabla f(\bar{x}) = 0. \quad (1.1.5)$$

Suppose now that  $f$  is twice differentiable.

- (b) Second-order necessary condition: if  $\bar{x}$  is a local minimum then

$$\langle h, \nabla^2 f(\bar{x})h \rangle \geq 0 \quad \text{for all } h \in \mathbb{R}^n. \quad (1.1.6)$$

- (c) Second-order sufficient condition: if  $\bar{x}$  satisfies (1.1.5) together with

$$\langle h, \nabla^2 f(\bar{x})h \rangle > 0 \quad \text{for all } h \in \mathbb{R}^n \setminus \{0\}, \quad (1.1.7)$$

then  $\bar{x}$  is a local minimum.

PROOF. Exercise; everything is based on the following developments: for  $h$  arbitrary in  $\mathbb{R}^n$ ,

$$\begin{aligned} f(\bar{x} + h) &= f(\bar{x}) + \langle \nabla f(\bar{x}), h \rangle + o(\|h\|) \\ &= f(\bar{x}) + \langle \nabla f(\bar{x}), h \rangle + \frac{1}{2} \langle h, \nabla^2 f(\bar{x})h \rangle + o(\|h\|^2). \end{aligned}$$
□

Among the above optimality conditions, (1.1.5) is the most used and there is a name for it:

**Definition 1.1.4** A *critical*, or *stationary* point for the differentiable  $f$  is an  $x \in \mathbb{R}^n$  satisfying  $\nabla f(x) = 0$ . □

A genuinely necessary and sufficient condition for local optimality does not exist, but observe that the difference between (1.1.6) and (1.1.7) is fairly small: the latter requires that all the eigenvalues of  $\nabla^2 f(\bar{x})$  be positive, while the former allows some of them to be zero. Anyway, filling the gap between (b) and (c) would amount to analyzing higher order expansions of  $f$ : for those  $h$  having  $\langle h, \nabla^2 f(\bar{x})h \rangle = 0$ , the third-order term must be 0 and the fourth-order term must be strictly positive (sufficient condition) or at least nonnegative (necessary condition) etc.

This is not generally considered a very fruitful pastime, and it has been a tradition to limit the study to second order, which is already difficult enough in practice: to check whether a given  $\bar{x}$  is a local minimum, one must first check whether the gradient is 0 at  $\bar{x}$ ; then one must compute the second derivatives, and then check if they form a positive (semi-)definite operator. This is not practical as soon as  $n$  becomes large, say beyond some hundreds, which is common for optimization problems.

- (iv) Knowing that checking local optimality is already a difficult task in practice, computing a local minimum is even worse. In other words, (1.1.4) is not a tractable problem yet, and one has to be even more modest in solving (1.1.2); in fact, an optimization problem like (1.1.2) is considered as numerically “solved” if one has found a critical point in the sense of Definition 1.1.4. Yet, finding such a point is in general possible by approximation only, i.e. one must construct a sequence  $\{x_k\}$  such that

$$x_k \rightarrow \bar{x} \quad \text{with } \bar{x} \text{ stationary}$$

or at least

$$\nabla f(x_k) \rightarrow 0 \quad \text{when } k \rightarrow \infty$$

or, if we are less and less demanding,

$$\liminf_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0. \tag{1.1.8}$$

Let us sum up: to solve an optimization problem like (1.1.2), one is usually content with a sequence  $\{x_k\}$  satisfying (1.1.8). Accordingly, the following terminology is generally used:

**Definition 1.1.5** A minimization algorithm is said to be *convergent*, or also *globally convergent*, for  $f$  in a given class (say a  $C^1$  function) if (1.1.8) holds for all  $x_1 \in \mathbb{R}^n$  and all  $f$  in this class. □

We mention here that the terminology “globally convergent” is misleading, because it does not mean convergence to a global minimum (1.1.1) (let us repeat that global optimality of a stationary point cannot be ascertained for a general  $f$ ). Here, the term “global” rather refers to the initial point  $x_1$ , which can be arbitrarily far from the cluster point  $\bar{x}$ .

## 1.2 General Structure of Optimization Schemes

To solve numerically our optimization problem, a set of rules must be defined which, knowing the objective  $f$  and starting from some initial  $x_1$ , construct the sequence  $\{x_k\}$  iteratively, in order to obtain (1.1.8).

In practice, of course,  $k$  does not go to infinity (otherwise no optimization problem would ever be solved!) so one of the things an algorithm must do at each iteration is answer the question: “Can the current  $x_k$  be considered an accurate enough approximation of some critical point?” (if yes stop; if no, proceed to computing  $x_{k+1}$ ). This is the *stopping criterion*, or stopping test, which directly conditions the time to be spent in solving the problem. According to our development above, the most natural stopping criterion is: stop if

$$\|\nabla f(x_k)\| \leq \delta \quad (1.2.1)$$

for some prescribed tolerance  $\delta > 0$ . If (1.1.8) holds, then (1.2.1) will certainly occur for some  $k$ ; when (1.2.1) occurs, some critical point can be hoped to exist close to  $x_k$  (although not necessarily, cf.  $f(x) = e^x$ ,  $x \in \mathbb{R}$ , which has no critical point although  $f'(x) = e^x$  can be arbitrarily close to 0).

**Remark 1.2.1** The above test (1.2.1) is not the only possibility for stopping an algorithm: a sound optimization process can and must contain several other tests; actually, designing good stopping criteria is not so simple and the question is not really solved in a totally satisfying way; occasionally, we will return to this later, especially in §3.

□

Thus we see that the complete optimization procedure, which is supposed to solve (1.1.1) or (1.1.5), is made up of several ingredients: the algorithm itself,  $f$ ,  $x_1$ ,  $\delta$ , etc (this list being non-exhaustive). They fall into two categories:

- (U) those pieces that characterize the problem to be solved; they are within the responsibility of the *user*, who has posed the problem and who is interested in knowing its solution; there we find: the choice of the initial  $x_1$  (the user may have some idea of a solution point), of the tolerance such as  $\delta$  for (1.2.1) (only the user knows how accurately the problem should be solved) and, last but not least, the objective function  $f$  itself;
- (A) the second category contains the algorithm proper, i.e. the set of rules to construct the iterative sequence  $\{x_k\}$ ; it is in the responsibility of the algorithm’s *designer*, who has defined the rules for iteration.

If we are dealing with *general* optimization – as opposed to a problem with special structure and a method especially tailored for it – the above two categories (U) and

(A) are fairly independent of each other. The most important part in (U), namely the definition of  $f$ , does not depend on the particular algorithm that is going to minimize it! On the other hand, and this is the most important point, the algorithm (the set of rules) is also totally independent of the actual  $f$ ; usually, it has in fact been designed long before the particular problem was posed and long before the optimization process is executed. We mention, however, that this independency property has some exceptions; for example, a tolerance such as  $\delta$  for (1.2.1) cannot be chosen totally *in abstracto*: there are algorithms whose convergence in the sense of (1.1.8) is so slow that, in practice, they cannot accommodate very small values of  $\delta$ .

In most cases, an optimization process is a computer program, or something similar. This means that the sequence of operations that make up the process are somehow automatized: they have been organized *before* the actual execution of the process, during which no human action can be taken. We will see later, particularly in Chap. XII, that some optimization problems (having *decentralized* character) can be made up of several programs, not coexisting in the same computer; worse, they may not even be in the same computing center. Under these conditions, the dichotomous structure (U) – (A) mentioned above must be reflected in the optimization set, the two parts of which must be clearly identified and separated. For the user, (A) is a *black box* which, when fed with  $x_1, \delta$  and  $f$ , outputs its last iterate  $x_K$ , approximately optimal if possible. Conversely, (U) is for the algorithm's designer another *black box*, actually made up of two parts: one part is the “driver”, or the “main program”, which sets up the problem, prepares the work for the algorithm and gives it some general instructions such as  $x_1, \delta$  etc; the second part contains the definition of  $f$  itself; when needed, it computes informations about  $f$  at a given  $x$ .

In this chapter, as well as later throughout this book, we assume that this second part in the user's black box (U) computes the value  $f(x)$  of  $f$  at a given  $x$ , and also the value  $\nabla f(x)$  of its gradient. We will always use the notation  $s$  for the gradient; so from now on we define

$$s(x) := \nabla f(x), \quad s_k := \nabla f(x_k) \quad \text{etc.}$$

**Remark 1.2.2** In most applications, the numerical value of  $\nabla f(x)$  can be calculated in a computing time that is of the same order as the time needed to calculate the numerical value of  $f(x)$ . On the other hand, it is quite usual in applications that the *human time* needed to compute the formal derivatives (i.e. to write the corresponding computer program) is much larger than for  $f$ : examples where it takes a man-year are not exceptional; computing the gradient can be quite an investment.  $\square$

To summarize, the optimization schemes that we will always consider are organized as illustrated in Fig. 1.2.1. Blocks (U0) and (U1) are the two parts in (U) mentioned above (the driver and the characterization of  $f$ ) and (A) is the algorithm. The optimization process starts with execution of (U0), until everything is ready to construct the iterations  $\{x_k\}$ ; then the control is passed to the black box (A) which, during the iterations, passes the control temporarily and regularly to the black box (U1) in order to get the necessary information concerning  $f$  for various  $x$ -values.

Our point of view will be that of the designer and our aim is to study optimization algorithms only, i.e. what is inside (A); we therefore simply assume that (U1) exists to compute  $f(x)$  and  $s(x)$  at any  $x$  that we may decide.

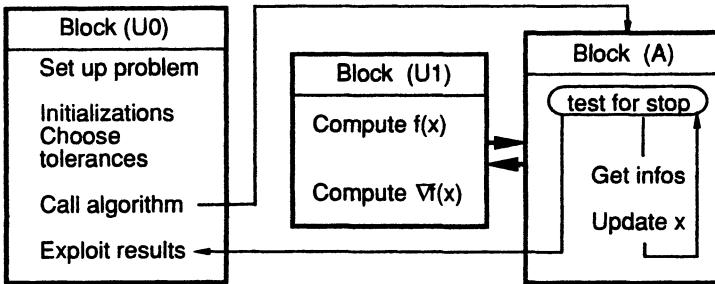


Fig. 1.2.1. General organization of an optimization program

**Remark 1.2.3** It is important to realize that the only information available from  $f$  (apart from general properties like its degree of smoothness) reduces to (U1), or rather to executions of (U1); it is a good idea to think of (U1) as, say, a computer tape, unreadable for a human being. The information available is, therefore, purely *pointwise*; it is not permissible, for example, to choose  $r \in \mathbb{R}$  and to say: “Let us take  $x$  such that  $f(x) = r$ ”, or even “... such that  $f(x) \leq r$ ”; although there may be many such  $x$ ’s, finding just one is already a nontrivial problem.

### 1.3 General Structure of Optimization Algorithms

We study now in more detail the construction of the sequence  $\{x_k\}$  mentioned in §1.1: how can it be a “good”, “minimizing” sequence, reaching (1.2.1) as soon as possible?

Most minimization algorithms are so-called *descent algorithms*, in the sense that  $f$  is forced to decrease at each iteration:

$$f(x_{k+1}) < f(x_k) \quad \text{for } k = 1, 2, \dots \quad (1.3.1)$$

In view of the limited information available from  $f$  (see the end of §1.2) it may be necessary to try several  $x$ -values before the actual move can be made from  $x_k$  to  $x_{k+1}$ . An iteration of a descent algorithm is essentially a *trial and error* process, roughly comparable to a walk in the dark toward the top of a hill:  $x \in \mathbb{R}^n$  is the position,  $-f(x)$  is the altitude to be maximized. The hiker does not have a direct, continuous, feeling of his altitude: he can only measure it at discrete moments, by a call to (U1), and it takes time to measure it. On the other hand, (U1) is more than an altimeter since it also gives the local variation of the altitude (the gradient of  $f$ ). At a given  $x$ , the hiker must estimate the behaviour of the terrain around him in order to guess where to go. Then he moves and checks whether he seems to get closer to his target. If not, he must retrace his steps and try again.

For most classical algorithms, an iteration starting from the current iterate  $x_k$  is composed of two stages.

- The first stage, the *direction-finding* procedure, consists of finding  $d_k \in \mathbb{R}^n$ , interpreted as a direction along which it is worth looking for the next iterate  $x_{k+1}$ . To

compute this direction, a local study around  $x_k$  is made, and the original problem is approximated by a simpler one providing an easy guess where to go.

- The second stage, called the *line-search*, is the actual computation of  $x_{k+1}$ : one computes a stepsize  $t_k > 0$  along  $d_k$ , and then  $x_k$  is updated to  $x_{k+1} = x_k + t_k d_k$ ; in contrast with the first stage, this computation is made upon observation of the *true original problem*, i.e. upon direct calls to (U1), so as to obtain exact values of  $f$ .

A classical optimization algorithm, therefore, presents itself schematically as follows:

**Algorithm 1.3.1 (Schematic Descent Algorithm)** The initial point  $x_1 \in \mathbb{R}^n$  and the tolerance  $\delta > 0$  are given, as well as the black box (U1) which computes  $f(x)$  and  $\nabla f(x)$  for arbitrary  $x \in \mathbb{R}^n$ . Set  $k = 1$ .

STEP 1 (Stopping criterion). If  $\|\nabla f(x_k)\| \leq \delta$  stop.

STEP 2 (Finding the direction). With the help of a model of the problem around  $x_k$ , find  $d_k \in \mathbb{R}^n$  for which

$$\exists t > 0 \quad \text{such that} \quad f(x_k + t d_k) < f(x_k).$$

STEP 3 (Line-search). By repeated calls to (U1) at  $x_k + t d_k$  for various values of  $t$ , find a convenient  $t_k > 0$ , satisfying in particular

$$f(x_k + t_k d_k) < f(x_k).$$

STEP 4 (Loop). Set  $x_{k+1} = x_k + t_k d_k$ , replace  $k$  by  $k + 1$  and loop to Step 1.  $\square$

**Remark 1.3.2** It is in Step 3 that the trial and error process mentioned at the beginning of this Section 1.3 takes place. The descent property (1.3.1) must be obtained there and, knowing from  $f$  only the local information contained in the black box (U1), it is not a trivial matter – recall Remark 1.2.3.  $\square$

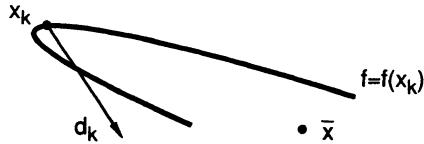
The methods functioning along this descent principle could be called *local methods*. It is a privilege of optimization to furnish a direct criterion (namely  $f$ ) measuring how good a given  $x$  can be. An important advantage of the above technique 1.3.1, in which the objective function  $f$  is improved at each  $k$ , is that it automatically ensures *stability*. Considering that optimization algorithms are after all methods to solve nonlinear equations like  $\nabla f(x) = 0$ , one may ask whether there is any difference between such algorithms and general equation-solvers. There is indeed a difference: to construct a sequence  $\{x_k\}$  for solving

$$F(x) = 0$$

where  $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is not a gradient, a very basic difficulty is to have  $\{x_k\}$  *bounded* (an essential requirement for  $x_k$  to converge to some solution  $\bar{x}$ !). In optimization, forcing  $f$  to decrease at each iteration tends to stabilize  $\{x_k\}$ , which in particular must lie in the sublevel-set  $S_f(x_1)(f)$ . In most applications, this sublevel-set is bounded. Said otherwise, an unbounded  $\{x_k\}$  would suggest some ill-posedness of the original optimization problem, rather than a failure of the algorithm.

The advantage of the descent property has its price: requiring  $f(x_k + t_k d_k) < f(x_k)$  can drastically restrict the move from  $x_k$  to  $x_{k+1}$ ; very often, it results in little progress toward the solution, unless the direction  $d_k$  is an excellent one and points inside regions where  $f$  is

largely decreasing. This is illustrated by Fig. 1.3.1, showing a sublevel-set of  $f$ : a long step along  $d_k$  would place  $x_{k+1}$  much closer to the solution  $\bar{x}$ , but this long step is forbidden if the sublevel-set is narrow and elongated. One may think that the situation considered in this picture is particularly unfavourable; it is actually quite common – in fact it is the rule as soon as  $x_k$  starts approaching  $\bar{x}$ . The lesson is that the direction must be chosen with great care.



**Fig. 1.3.1.** Descent methods can be slow

In view of this division of one iteration into two stages, we will study successively the direction finding, and then the line-search.

## 2 Defining the Direction

### 2.1 Descent and Steepest-Descent Directions

The next definition is motivated by our care to decrease  $f$  at each iteration: we want

$$\exists t > 0 \quad \text{such that} \quad f(x_k + td) < f(x_k). \quad (2.1.1)$$

□

**Definition 2.1.1** A *descent direction* issued from  $x$  for the continuously differentiable  $f$  is a  $d \in \mathbb{R}^n$  such that (recall our notation  $s = \nabla f$ )

$$\langle s(x), d \rangle < 0. \quad (2.1.2)$$

□

Clearly, if  $f$  is a fixed given function, there may exist directions which satisfy the natural property (2.1.1) but not (2.1.2) – think of  $f(x) := -\|x\|^2$  at  $x = 0$ : every  $d \neq 0$  satisfies (2.1.1) and no  $d$  satisfies (2.1.2). Definition 2.1.1 then appears somewhat artificial. However, one should remember Remark 1.2.3 and the rules of the game for numerical optimization: if  $f$  is an *arbitrary* function compatible with the known information  $f(x)$  and  $s(x)$  at a given fixed  $x$ , then (2.1.2) is the only chance to obtain (2.1.1):

**Proposition 2.1.2** Let the triple  $\{x_k, f_k, s_k\}$  be given in  $\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^n$  and consider the set of functions

$$\Phi_k := \{f \text{ differentiable at } x_k : f(x_k) = f_k, \nabla f(x_k) = s_k\}.$$

Then,  $d \in \mathbb{R}^n$  satisfies (2.1.1) for any  $f \in \Phi_k$  if and only if  $\langle s_k, d \rangle < 0$ .

PROOF. [if] Take  $d$  with  $\langle s_k, d \rangle < 0$  and  $f$  arbitrary in  $\Phi_k$ ; then

$$f(x_k + td) = f(x_k) + t\langle s_k, d \rangle + o(t)$$

and it suffices to take  $t > 0$  small enough to obtain (2.1.1).

[only if] Take  $d$  with  $\langle s_k, d \rangle \geq 0$  and  $\hat{f} \in \Phi_k$  defined by

$$\hat{f}(\cdot) := f_k + \langle s_k, \cdot - x_k \rangle;$$

then, there holds

$$\forall t \geq 0 \quad \hat{f}(x_k + td) = \hat{f}(x_k) + t\langle s_k, d \rangle \geq \hat{f}(x_k),$$

so this  $d$  cannot satisfy (2.1.1) for all  $f \in \Phi_k$ .  $\square$

Thus, Definition 2.1.1 does appear as *the* relevant concept for a descent direction. It implies more than (2.1.1), namely

$$\exists \bar{t} > 0 \quad \text{such that} \quad f(x + td) < f(x) \quad \text{for all } t \in ]0, \bar{t}]$$

and this makes sense since, in optimization algorithms, the move from  $x_k$  to  $x_{k+1}$  – and hence  $t_k$  – is usually quite small (remember Fig. 1.3.1).

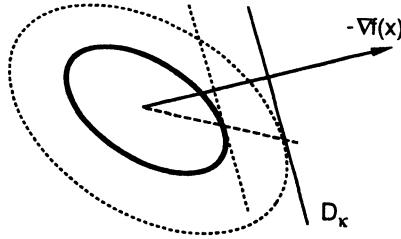
A descent direction in the sense of Definition 2.1.1 is one along which not only does  $f$  decrease, but it does so at a non-negligible rate, i.e. the decrease in  $f$  is proportional to the move from  $x$ . This rate of decrease, precisely, is the number  $\langle s(x), d \rangle$ , the *directional derivative* of  $f$  at  $x$  in the direction  $d$  (see Remark I.4.1.4). It is the derivative at 0 of the univariate function  $t \mapsto f(x + td)$  and it measures the above-mentioned progress that is made locally when moving away from  $x$  in the direction  $d$ . Then it is a natural idea to choose  $d$  so as to make this number as negative as possible, a concept which we now make precise:

**Definition 2.1.3** Let  $\|\cdot\|$  be a norm on  $\mathbb{R}^n$ . A *normalized steepest-descent direction* of  $f$  at  $x$ , associated with  $\|\cdot\|$ , is a solution of the problem

$$\min \{\langle s(x), d \rangle : \|d\| = 1\}. \quad (2.1.3)$$

A non-normalized steepest-descent direction is a  $d \neq 0$  such that  $\|d\|^{-1}d$  is a normalized steepest-descent direction.  $\square$

Problem (2.1.3) does have optimal solutions because the (continuous) function  $\langle s(x), \cdot \rangle$  attains its minimum on the (compact) boundary of the unit ball; it may have several solutions (see §2.2 below). To characterize these solutions, the results of Chap. VII are needed. For our present purpose, however, it suffices to display them graphically, which is done on Fig. 2.1.1: for given  $\kappa \in \mathbb{R}$ , the locus of those  $d$  having  $\langle s(x), d \rangle = \kappa$  is an affine hyperplane  $D_\kappa$  orthogonal to  $s(x)$ ; the optimal solutions are obtained for  $\kappa$  as small as possible, i.e. when  $D_\kappa$  is as far as possible in the direction of  $-s(x)$ , yet touching the unit ball.



**Fig. 2.1.1.** Homothety in the steepest-descent problem

**Remark 2.1.4** Figure 2.1.1 displays the need for a normalization in Definition 2.1.3: without its constraint, problem (2.1.3) would have no solution (or, rather, a solution “at infinity”, with a directional derivative “equal to  $-\infty$ ”) and the concept would not make sense. A norm  $\|\cdot\|$ , which does not have to be the Euclidean norm  $\|\cdot\| = \langle \cdot, \cdot \rangle^{1/2}$ , must be specified when speaking of a steepest-descent direction.

This implies also the artificial introduction of the number 1 in (2.1.3). It should be noted, however, that the particular value “1” is irrelevant, as far as a steepest-descent *direction* is of interest, regardless of its length. Looking at Fig. 2.1.1 with different glasses, we observe that collinear solutions are obtained if,  $\kappa$  being kept fixed, say  $\kappa = -1$ , the radius of the unit ball is changed so as to become as small as possible yet touching  $D_{-1}$ . In other words, (2.1.3) and

$$\min \{ \|d\| : \langle s(x), d \rangle = -1 \} \quad (2.1.4)$$

have collinear solutions. This property is due to homothety in Fig. 2.1.1: the functions  $\langle s(x), \cdot \rangle$  and  $\|\cdot\|$  are positively homogeneous of degree 1. This remark explains the important property that replacing “1” by  $\kappa > 0$  in (2.1.3) or (2.1.4) would just multiply the set of optimal solutions by  $\kappa$ . Within a descent algorithm, this multiplication would be cancelled out by the line-search and, finally, the only important definition for our purpose is that of non-normalized (steepest-descent) directions.  $\square$

The choice of the norm in (2.1.3) or (2.1.4) is of fundamental importance for practical efficiency, and we will divide our study into two parts, according to this choice. Afterwards, we will study the conjugate-gradient method, which is based on a different principle.

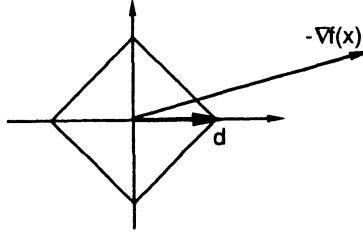
## 2.2 First-Order Methods

The first possibility for the norm in (2.1.3) is a choice *a priori*, independent of  $f$ . Classically, there are two such choices: the  $\ell_1$  norm and the Euclidean norm.

**(a) One Coordinate at a Time** The  $\ell_1$  norm is

$$\|d\| = |d|_1 := \sum_{i=1}^n |d^i| \quad (2.2.1)$$

(here and in what follows,  $\mathbb{R}^n$  is assumed to have a basis, in which  $z^i$  is the  $i^{\text{th}}$  coordinate of a vector  $z$ , and the natural dot-product is used). Figure 2.2.1 particularizes Fig. 2.1.1



**Fig. 2.2.1.** An  $\ell_1$ -steepest-descent direction

to this norm. It clearly indicates the following characterization of an optimal  $d_k$  (which will be confirmed in Chap. VII): let  $i_k$  be an index such that

$$|s^{i_k}(x_k)| \geq |s^i(x_k)| \quad \text{for } i = 1, \dots, n$$

(note that there may be several such  $i_k$ , just choose one); then the  $n$  numbers

$$d_k^i = \begin{cases} 0 & \text{if } i \neq i_k, \\ -\frac{s^{i_k}(x_k)}{|s^{i_k}(x_k)|} & \text{if } i = i_k \end{cases} \quad (2.2.2)$$

make up an optimal direction. In other words: among the solutions of (2.1.3) with the  $\ell_1$  norm (2.2.1), there is one of the vectors of the basis of  $\mathbb{R}^n$  (neglecting its sign, chosen so as to obtain a descent direction), namely one corresponding to a maximal coordinate of the gradient.

**Remark 2.2.1** Under these conditions,  $x_{k+1}$  is obtained from  $x_k$  by changing only one coordinate, namely one which locally changes  $f$  most. The resulting scheme has an interpretation in terms of a traditional method, the method of Gauss-Seidel, which we briefly describe now. To solve the linear system

$$Qx + b = 0, \quad (2.2.3)$$

this method consists of choosing at iteration  $k$  one of the equations, say the  $i_k^{\text{th}}$ , and of solving this equation with respect to the single variable  $x^{i_k}$ , the other “unknowns”  $x^j$  being set to the (known) coordinates of the current  $x_k$ . In other words, the whole vector  $x_{k+1}$  is just  $x_k$ , except for its  $i_k^{\text{th}}$  coordinate, which is set to the value  $\alpha \in \mathbb{R}$  solving

$$\sum_{j \neq i_k} q_{i_k j} x_k^j + q_{i_k i_k} \alpha + b_{i_k} = 0 \quad (2.2.4)$$

(the method is well-defined if all diagonal entries of  $Q$  are nonzero).

Now, suppose that our function to be minimized

$$f(x) = \frac{1}{2} \langle Qx, x \rangle + \langle b, x \rangle$$

is quadratic with  $Q$  symmetric positive definite; the gradient  $s(x) = Qx + b$  is affine and minimizing  $f$  is just solving (2.2.3). Then, the Gauss-Seidel iterate given by (2.2.4) has the form  $x_k + t_k d_k$  with  $d_k$  given by (2.2.2). It can be shown also that the stepsize  $t_k$  corresponding to  $\alpha$  is positive and actually minimizes  $f$  along  $d_k$  (this is due to positive definiteness of  $Q$ , which implies in particular  $q_{i_k i_k} > 0$ ). In other words, the  $\ell_1$ -steepest-descent method is simply a variant of the method of Gauss-Seidel, in which:

- $i_k$  is an index giving a most violated equality in (2.2.3) (in the original method,  $i_k$  was rather chosen cyclically, thus resulting in a direction totally blind to the behaviour of  $f$ ),
- there is some freedom for the actual value of  $\alpha$ , obtainable by some line-search instead of as a solution of (2.2.4).

**(b) Euclidean Steepest Descent** The second classical choice for  $\|\cdot\|$  in (2.1.3) is simply the Euclidean norm  $\|\cdot\| := \|\cdot\|$  induced by the scalar product  $\langle \cdot, \cdot \rangle$  defining  $\nabla f$ . When Fig. 2.1.1 is particularized to this case, the spheres  $\|d\| = \kappa$  become “ordinary” circles and it is easy to realize that  $d_k$  is then  $-s(x_k)$  (up to the normalization coefficient). We therefore obtain:

**Definition 2.2.2** The *gradient* method is the steepest-descent method in which the norm  $\|\cdot\|$  for Definition 2.1.3 is  $\langle \cdot, \cdot \rangle^{1/2}$ . In this method, the next iterate is looked for in the form

$$x_{k+1} = x_k - t \nabla f(x_k),$$

the stepsize  $t > 0$  being given by a line-search.  $\square$

**Remark 2.2.3** Just as in Remark 2.2.1, we also have an interpretation of the gradient method. To solve a system of equations (linear or not)

$$s(x) = 0, \quad (2.2.5)$$

where  $s : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is or is not a gradient mapping, a classical method consists of defining the sequence of iterates by

$$x_{k+1} = x_k + \rho s(x_k) \quad (2.2.6)$$

with a suitable parameter  $\rho \in \mathbb{R}$ . All the coordinates of the current iterate are modified at the same time, instead of one by one as in the case of Gauss-Seidel. A motivation to do so is that (2.2.5) is equivalent to

$$x = x + \rho s(x)$$

where  $\rho \neq 0$  is arbitrary; (2.2.6) is then the first idea that comes to mind, namely the process of successive approximations, to solve the above fixed-point problem. It remains to choose  $\rho$  and we have here an illustration of Remark 1.3.2. To solve (2.2.5) by (2.2.6), the question of choosing a suitable value for  $\rho$  is in fact puzzling – for convergence as well as numerical efficiency. Here however, knowing that  $s$  is actually the gradient of some function  $f$  which must be *minimized*, and being able to actually calculate  $f$ , provides decisive information: (i) it indicates that  $\rho (= -t)$  should be negative, and (ii) it gives a constructive way of adapting  $\rho$  at each iteration, via a line-search.

**(c) General Normings** The above two steepest-descent methods (Gauss-Seidel and gradient) are only two instances of an infinite number of possibilities, each corresponding to a particular  $\|\cdot\|$  in (2.1.3). Among all these norms, it is natural to ask whether there is a “best” one, yielding a “best” method in some sense? Whatever “best” means, the norm in question should be “universal”, i.e. chosen *a priori* and independent of the particular  $f$  to be minimized (by contrast, §2.3 will be devoted to a norm depending on  $f$  and on the iteration index  $k$ ).

In the context of an algorithm of the type 1.3.1, an essential characteristic of the direction  $d_k$  is the angle that it makes with the gradient; thus, for nonzero  $s$  and  $d$  in  $\mathbb{R}^n$ , set

$$\cos(s, d) := \frac{\langle s, d \rangle}{\|s\| \|d\|}.$$

By the Cauchy-Schwarz inequality, any value of  $\cos$  is in  $[-1, +1]$ . To say that  $d$  is a descent direction at  $x$  is to say that  $\cos(\nabla f(x), d) \in [-1, 0[$ ; in the gradient method,  $\cos(s_k, d_k) = -1$  is as negative as possible.

It so happens that the choice of  $\|\cdot\|$  has little influence on whether  $x_k$  converges or not to a critical point (for this, the line-search is much more crucial). On the other hand, the value of  $\cos(s_k, d_k)$  does influence the *speed* of that convergence, i.e. how fast (1.2.1) can be obtained. The following result gives a useful indication in this line; we omit the proof, which is of little interest in the framework of this book.

**Theorem 2.2.4** *Let  $\{x_k\}$  be the sequence generated by a steepest-descent method in which  $t_k$  is chosen as a solution of  $\min_{t>0} f(x_k + t d_k)$ . Suppose that  $f$  is twice continuously differentiable, that  $x_k \rightarrow \bar{x}$  with  $\nabla f(\bar{x}) = 0$ , and that there are two constants  $0 < \ell \leq L$  such that, for all  $h \in \mathbb{R}^n$  and  $x$  close enough to  $\bar{x}$ ,*

$$\ell \|h\|^2 \leq \langle \nabla^2 f(x) h, h \rangle \leq L \|h\|^2. \quad (2.2.7)$$

*Suppose also that there is  $C \in ]0, 1]$  such that*

$$\cos(s_k, d_k) \leq -C \quad \text{for } k = 1, 2, \dots \quad (2.2.8)$$

*Then there is  $M > 0$  such that, for  $k$  large enough:*

$$|f(x_k) - f(\bar{x})| \leq M [1 - (C\ell/L)^2]^k. \quad (2.2.9)$$

□

This result suggests that the error  $f(x_k) - f(\bar{x})$ , assumed to converge to 0, probably behaves like a geometric series with ratio close to 1 when  $C$  is close to 0. For given  $f$  ( $\ell/L$  fixed), a steepest-descent method with  $C \simeq 0$  (direction and gradient almost orthogonal) can be expected to converge much more slowly than the gradient method ( $C = 1$ ) – at least if the majorization in (2.2.9) is reasonably sharp.

Consider for example the method of Gauss-Seidel and the usual dot-product  $\langle s, d \rangle = s^\top d$  in  $\mathbb{R}^n$ . It is easy to see that, with  $d_k$  of (2.2.2), there holds

$$\langle s_k, d_k \rangle = -|s^{i_k}(x_k)|, \quad \|d_k\| = 1 \quad \text{and} \quad \|s_k\|^2 \leq n |s^{i_k}(x_k)|^2$$

so (2.2.8) holds with  $C = 1/\sqrt{n}$ . If, once again, the majorations (2.2.8) and (2.2.9) are reasonably sharp, the method of Gauss-Seidel becomes drastically slow when  $n$  becomes large. For the gradient method, however,  $C = 1$  is independent of  $n$  and the rate of convergence does not deteriorate when  $n \rightarrow \infty$ .

**Example 2.2.5** To illustrate the above comments, take the symmetric positive definite  $n \times n$  matrix  $Q$  defined by

$$Q_{ij} = \begin{cases} 2 & \text{if } i = j \\ -1 & \text{if } |i - j| = 1 \\ 0 & \text{if } |i - j| > 1 \end{cases}$$

(which is common when differential equations are discretized) and consider the problem of minimizing

$$f(x) := \frac{1}{2}\|x\|^2 + \frac{1}{2}\langle Qx, x \rangle \quad \text{for } x \in \mathbb{R}^n.$$

As a quadratic function,  $f$  has the constant Hessian  $I + Q$ , for which (2.2.7) holds with  $\ell/L \simeq 0.2$ . Therefore, the gradient method should converge roughly as  $(0.96)^k$  (at least), independently of  $n$ . On the other hand, the method of Gauss-Seidel might have the rate of convergence  $1 - 0.04/n$ .

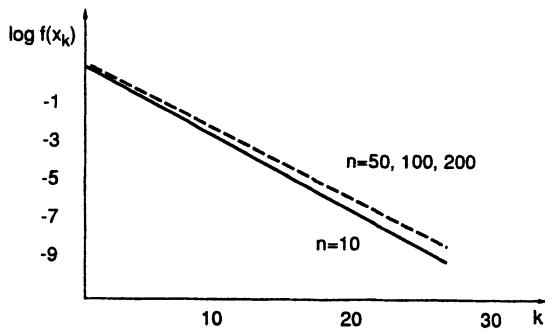


Fig. 2.2.2. Typical behaviour of gradient method

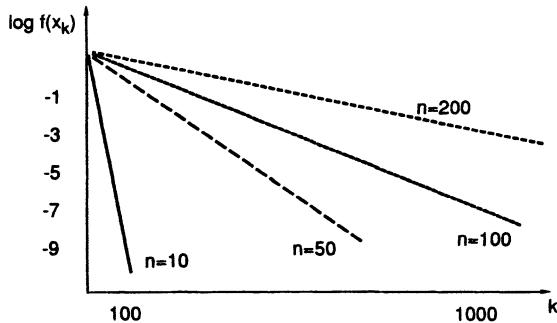


Fig. 2.2.3. Corresponding behaviour of Gauss-Seidel

This is confirmed graphically: Fig. 2.2.2 displays, for various values of the dimension  $n$ , the decrease of  $\log f(x_k)$  obtained by the gradient method, as a function of the iteration-number  $k$ . It is remarkably constant and convergence can be considered as obtained in some 20–30 iterations. On the other hand, Fig. 2.2.3 displays the decrease obtained with the method of Gauss-Seidel, which clearly deteriorates for large  $n$ . For  $n \geq 200$ , it becomes so slow that the method must be considered as non-convergent (note the difference in horizontal scales between Figures 2.2.2 and 2.2.3!).  $\square$

**Remark 2.2.6** The gradient method thus appears as optimal in some sense among all the steepest-descent methods, since it minimizes  $\cos(s_k, d_k)$  at each iteration. This property should not be misinterpreted, however: there may well be functions for which Gauss-Seidel's method is much faster (think of  $f(\xi^1, \dots, \xi^n) = \sum_i (i\xi^i)^2$ , for example!). The gradient method is optimal in a “minimax” sense, among the functions satisfying the assumptions of Theorem 2.2.4.  $\square$

Let us conclude this section by a terse comment: all these first-order methods should **NEVER** be used, because they are highly inefficient in terms of speed of convergence. One can just say that there is some excuse for using the  $\ell_1$  method: it does not really require computing derivatives (if the choice for  $i_k$  in (2.2.2) is cyclic instead of optimal and if the line-search accepts negative stepsizes); on the other hand, it is admittedly even worse than the gradient method (recall Example 2.2.5). Actually, the methods of the next sections go so much faster, and with so little additional cost, that it would be a sin against economy to give preference to the present first-order methods.

### 2.3 Newtonian Methods

The methods of §2.2 are of first-order type in the sense that the objective function  $\langle \nabla f(x), \cdot \rangle$  of (2.1.3) is a first-order model for  $f$ . By contrast, the methods of the present section build a second-order model, which is then used in the normalization constraint of (2.1.3) to improve the direction.

Let  $Q$  be a symmetric positive definite operator; then  $\langle Qx, x \rangle$  defines (the square of) a norm; (2.1.3) can therefore be specialized to

$$\min \{ \langle s(x_k), d \rangle : \langle Qd, d \rangle = 1 \} \quad (2.3.1)$$

(note that the presence or absence of the square root in the constraint does not matter a bit). It so happens that (2.3.1) has a unique solution which is collinear to that of

$$\min_d [\langle s(x_k), d \rangle + \frac{1}{2} \langle Qd, d \rangle] \quad (2.3.2)$$

(this will be confirmed in Chap. VII). Now, if  $\nabla^2 f(x_k)$  happens to be positive definite and if we take  $Q := \nabla^2 f(x_k)$ , we obtain a method with a sensible rationale: (2.3.2) then consists of minimizing the second-order development of  $f$  near  $x_k$

$$\min_d [\langle \nabla f(x_k), d \rangle + \frac{1}{2} \langle \nabla^2 f(x_k) d, d \rangle]. \quad (2.3.3)$$

**Remark 2.3.1** Thus, the direction can now be computed in three different ways, which are essentially equivalent: by solving (2.3.1) = (2.1.3), by solving (2.1.4), or by solving (2.3.2). These three ways only differ by the *length* of the resulting direction, this length being directly related to the value “1” in (2.3.1) and (2.1.4). When  $Q = \nabla^2 f(x_k)$ , the form (2.3.2) = (2.3.3) is superior to the other two because it gives not only a direction but also a length, i.e. a *stepsize*.

Clearly enough, (2.3.3) is of particular interest since its solution minimizes the second-order approximation of  $d \mapsto f(x_k + d)$ . On the other hand, if  $d_k$  solves (2.3.3) and if  $\kappa > 0$ , then  $\kappa d_k$  minimizes (remember Remark 2.1.4)

$$\langle \nabla f(x_k), d \rangle + \frac{1}{2\kappa} \langle \nabla^2 f(x_k) d, d \rangle$$

which has little to do with a second-order approximation of  $f$ . As a result, the stepsize  $t = 1$ , which yields the iterate  $x_k + d_k$ , is supposedly better than any other stepsize. By contrast, neither (2.1.3) nor (2.1.4) gives an idea of what the stepsize should be. This will have some consequences for the line-search.  $\square$

**Remark 2.3.2** As in §2.2, we can draw a parallel between the present second-order method and equation-solving: consider again a system like  $s(x) = 0$ . Apart from the process of successive approximations, the next idea for solving it is Newton's method. Starting from the current iterate  $x_k$ , we would like the next iterate  $x_k + d$  to solve

$$s(x_k + d) = 0.$$

To mimic this, we replace  $s$  locally around  $x_k$  by its first-order approximation and we solve instead

$$[s(x_k + d) \simeq] \quad s(x_k) + Js(x_k)d = 0,$$

where  $Js$  is the Jacobian operator of  $s$ . Here, in the context of optimization, the Jacobian operator of  $s := \nabla f$  is the Hessian operator of  $f$  and the above linear system is

$$[\nabla f(x_k + d) \simeq] \quad \nabla f(x_k) + \nabla^2 f(x_k)d = 0, \quad (2.3.4)$$

which is nothing other than the optimality condition for (2.3.3). All this is perfectly normal: a second-order development of  $f$  corresponds to a first-order development of  $\nabla f$ .

The qualities and deficiencies of Newton's method are well-known:

- (i) it is an extremely fast method, with so-called 2<sup>nd</sup> order Q-convergence. Roughly speaking, this means that if the current iterate has  $\ell$  exact digits, the next iterate has  $2\ell$  exact digits; *but*
- (ii) it often diverges violently, especially if  $x_1$  is not close to the solution of  $s(x) = 0$ ;
- (iii) it requires computing second derivatives – which is usually a highly unpleasant task for the user – and then solving a linear system such as (2.3.4); all this is somewhat heavy and may not be convenient.
- (iv) Another disadvantage, peculiar to optimization problems, is that  $\nabla^2 f(x_k)$  must be positive definite, otherwise  $d_k$  may not be a descent direction. Note also that, if  $\nabla^2 f(x_k)$  is indefinite, (2.3.3) has usually no solution and (2.3.4) does not make sense for minimization: its solution(s) tend(s) to approximate a saddle-point or a maximum of  $f$ , not a minimum. Yet, while  $\nabla^2 f$  is normally positive definite in a neighborhood of a (local) minimum (if it varies continuously with  $x$  and if Theorem 1.1.3(c) applies), it has certainly no reason to be so in the whole space.

The aim of *Newtonian* methods is to eliminate (ii) – (iv) without destroying the advantage (i). Eliminating (ii) is quite easy and the line-search technique is made just for that (once again, recall Remark 1.3.2), provided that  $d_k$  is a descent direction.

**Remark 2.3.3** With Remark 2.3.1 in mind, the line-search should be understood as a mere safeguard against divergence of Newton's method, rather than a means to decrease  $f$  as much as possible – recall also the discussion following Remark 1.3.2. In other words, the stepsize  $t = 1$  should definitely be preferred, and only in case of total failure, whatever this means, should it be relinquished for another value.  $\square$

As for (iii) and (iv), they will be eliminated by the same single mechanism. The problem is that one does not want to compute  $\nabla^2 f$ , which in addition may not be convenient to use if it is not positive definite. So then the idea is: why not approximate it? Even more: why not approximate its inverse, which is in fact more useful for (2.3.4)? A further idea, related to (iv): the qualities of Newton's method are *local* only and  $\nabla^2 f(x_k)$  (or its inverse) is not much needed when  $x_k$  is far from a minimum. On the other hand,  $\nabla^2 f(x_k)$  usually becomes positive definite when  $x_k$  approaches a minimum. This provides two incentives for the following idea: why not take a positive definite approximation, albeit rough during the early stages of the descent process, when  $x_k$  is still far from a minimum?

With this last argument in mind, we realize that our problem resembles that of finding a minimum of  $f$ : what we need is another *algorithm* which, starting from some positive definite operator  $W_1$  and working in parallel with the construction of  $\{x_k\}$ , accumulates second-order information from  $f$  to construct a sequence  $\{W_k\}$  of positive definite operators. *Only at the end* of the process, need these operators approach the desired  $W^*$ , whatever it is.

Neglecting for the moment the issue of defining  $W^*$  (it should be something like  $[\nabla^2 f(\bar{x})]^{-1}$ , if  $x_k \rightarrow \bar{x}$ ), we obtain a specification of Algorithm 1.3.1:

**Algorithm 2.3.4 (Schematic Variable Metric Algorithm)** The initial point  $x_1 \in \mathbb{R}^n$  and the tolerance  $\delta > 0$  are given. Choose an initial symmetric positive definite operator  $W_1$ . Set  $k = 1$ ;  $s_k$  will denote  $\nabla f(x_k)$  as usual.

STEP 1 (Stopping criterion). If  $\|s_k\| \leq \delta$  stop.

STEP 2 (Finding the direction). Compute  $d_k = -W_k s_k$ .

STEP 3 (Line-search). Find  $t_k > 0$  and the corresponding  $x_{k+1} = x_k + t_k d_k$  satisfying in particular  $f(x_{k+1}) < f(x_k)$ .

STEP 4 (Metric update and loop). Select a new symmetric positive definite operator  $W_{k+1}$ ; replace  $k$  by  $k + 1$  and loop to Step 1.  $\square$

It remains to specify the choice of  $W_{k+1}$  in Step 4, the only new ingredient with respect to the general scheme 1.3.1. As seen already, the first property to be satisfied is:

$$W_{k+1} \text{ is symmetric positive definite for } k = 1, 2, \dots \quad (2.3.5)$$

Now, remembering that  $W_k$  is supposed to approach some inverse Hessian, we impose the following relation:

$$W_{k+1}(s_{k+1} - s_k) = x_{k+1} - x_k \quad (2.3.6)$$

known as the *secant equation*. Observe that  $x_{k+1}$  is already known in Step 4 of Algorithm 2.3.4, so  $s_{k+1}$  can be obtained from a call to block (U1) of Fig. 1.2.1; solving (2.3.6) for  $W_{k+1}$  is not an impossible task.

We will call *secant*, or *quasi-Newton* methods, the methods of the type 2.3.4 where  $W_k$  is updated in Step 4 so as to satisfy (2.3.6). The motivation for (2.3.6) can be explained as follows: making the necessary assumptions on  $f$ , start from the equality

$$\nabla f(x_{k+1}) - \nabla f(x_k) = \int_0^1 \nabla^2 f(x_k + t(x_{k+1} - x_k))(x_{k+1} - x_k) dt$$

and call

$$G_k := \int_0^1 \nabla^2 f(x_k + t(x_{k+1} - x_k)) dt$$

the mean value of  $\nabla^2 f$  between  $x_k$  and  $x_{k+1}$ . Then there holds by definition

$$G_k^{-1}[\nabla f(x_{k+1}) - \nabla f(x_k)] = x_{k+1} - x_k,$$

so (2.3.6) is a natural requirement since we would like  $W_k$  to resemble  $G_k^{-1}$ , at least asymptotically.

**Remark 2.3.5** The reason why (2.3.6) is called the secant equation is that it extends to the multi-dimensional case a known method to solve equations in one variable. Let the equation

$$s(x) = 0 \quad (s : \mathbb{R} \rightarrow \mathbb{R})$$

be solved by the one-dimensional Newton method ( $x_+$  is the next iterate,  $x$  is the current one)

$$x_+ = x - s(x)/s'(x).$$

This Newton method could be called the *tangent* method, as it approximates the graph of  $s$  by its tangent at the current  $x$ . Suppose that, just as in the multi-dimensional case, one does not wish to compute the derivative  $s'$ . Then, the idea of the *secant* method is to replace the above tangent by the secant crossing the graph of  $s$  at the two points  $(x, s(x))$  and  $(x_-, s(x_-))$  in  $\mathbb{R}^2$  ( $x_-$  is the previous iterate). Straightforward calculations show that the next iterate is now

$$x_+ = x - W s(x)$$

where we have set

$$W := \frac{x - x_-}{s(x) - s(x_-)}$$

and we recognize here the secant equation (2.3.6). □

The above remark shows that, if  $n = 1$ , there is exactly one quasi-Newton method – which, incidentally, need not satisfy (2.3.5). For  $n > 1$ , it is rather clear that the  $n$  scalar equations in (2.3.6) do not suffice to determine uniquely the  $\frac{1}{2}n(n+1)$  unknowns in  $W_{k+1}$ . In order to choose among the solutions, an additional requirement is imposed:

$$W_{k+1} \text{ should be "close" to } W_k \tag{2.3.7}$$

for rather understandable stability reasons: if  $\{W_k\}$  bounces back and forth between several values,  $\{d_k\}$  will do the same and  $\{x_k\}$  will oscillate, resulting in a loss of efficiency. Depending on the meaning chosen for (2.3.7), one obtains a potentially infinite number of variants.

The most widely used, known as BFGS (for C. Broyden, R. Fletcher, D. Goldfarb and D. Shanno), consists of taking  $W_{k+1}$  as follows: set

$$\xi = \xi_k := x_{k+1} - x_k \quad \text{and} \quad \sigma = \sigma_k := s_{k+1} - s_k; \tag{2.3.8}$$

then  $W_{k+1}$  is the operator which, to  $z \in \mathbb{R}^n$ , associates the image (we drop the subscript from  $W_k$  to alleviate notation)

$$W_{k+1}z = Wz + \left[ 1 + \frac{\langle \sigma, W\sigma \rangle}{\langle \sigma, \xi \rangle} \right] \frac{\langle z, \xi \rangle}{\langle \sigma, \xi \rangle} \xi - \frac{\langle \sigma, Wz \rangle \xi + \langle z, \xi \rangle W\sigma}{\langle \sigma, \xi \rangle}. \quad (2.3.9)$$

When  $\langle \cdot, \cdot \rangle$  is the usual dot-product,  $W_{k+1}$  has the explicit matrix expression

$$W_{k+1} = W + \frac{1 + \sigma^\top W \sigma \sigma^\top \xi}{\sigma^\top \xi} \xi \xi^\top - \frac{1}{\sigma^\top \xi} [\xi \sigma^\top W + W \sigma \xi^\top]. \quad (2.3.10)$$

Note: an expression like  $\xi \sigma^\top$  means the column matrix  $\xi$  post-multiplied by the row matrix  $\sigma^\top$ . The result is an  $n \times n$  matrix, whose kernel is the  $(n - 1)$ -dimensional subspace orthogonal to  $\xi$ , when  $\xi \neq 0$ .

A study of the convergence properties of secant methods would be far beyond the scope of this book. In relation with (2.3.5), we just mention without proof the following (important) result, which will be of interest to us later.

**Theorem 2.3.6** Suppose  $W_k$  is positive definite. A necessary and sufficient condition for  $W_{k+1}$  defined by (2.3.9) to be positive definite is  $\langle \sigma, \xi \rangle > 0$ .  $\square$

A last remark concerning actual implementations: in the case of the dot-product, when  $\{W_k\}$  is defined by (2.3.10), some prefer not to use  $W_k$  explicitly but rather its inverse (hence the approximation of the Hessian itself) in the product-form

$$M_k := W_k^{-1} = LDL^\top. \quad (2.3.11)$$

Here  $D$  is a positive diagonal matrix and  $L$  is a lower triangular matrix with diagonal elements all equal to 1. The system

$$M_k d + s_k = 0$$

is then solved via 3 easy systems:

$$Ly + s_k = 0, \quad Dz = y, \quad L^\top d = z.$$

## 2.4 Conjugate-Gradient Methods

The Newtonian methods of §2.3 are extremely efficient but become impractical when the number  $n$  of variables is large. Then it is impossible to store the quasi-Newton matrix, and furthermore the algebraic calculations may become overwhelming, as computing the direction involves  $O(n^2)$  operations. In this section, we introduce the *conjugate-gradient* methods, which have opposite characteristics. They use little storage, at the price of more modest speed of convergence, although they usually do not suffer the unacceptable behaviour of first-order methods of §2.2.

**(a) Linear Case** The rationale for conjugate-gradient methods is purely algebraic. It does not rely on a second-order development of the objective function, which is assumed to be *exactly* quadratic. Therefore suppose

$$f(x) := \frac{1}{2} \langle Qx, x \rangle + \langle b, x \rangle + c \quad (2.4.1)$$

is our quadratic objective function, where  $Q$  is a symmetric positive definite operator,  $b$  is an arbitrary vector in  $\mathbb{R}^n$  and  $c \in \mathbb{R}$ .

**Definition 2.4.1** At iteration  $k$  of a descent algorithm, call

$$U_k = \text{lin} \{s_1, \dots, s_k\} := \left\{ d \in \mathbb{R}^n : d = \sum_{i=1}^k \alpha_i s_i, (\alpha_1, \dots, \alpha_k) \in \mathbb{R}^k \right\}$$

the subspace spanned by the “accumulated information”  $s_1, \dots, s_k$ . The *conjugate-gradient method* to minimize  $f$  of (2.4.1) is the method in which, for each  $k$ ,  $x_{k+1}$  minimizes  $f$  in the affine manifold

$$V_k := \{x_k\} + U_k$$

parallel to  $U_k$  and containing  $x_k$ . □

Thus, our conjugate-gradient method minimizes  $f$  at the first iteration in the direction of the gradient, at the second iteration in the plane generated by  $s_1$  and  $s_2$  and passing by  $x_2$ , and so on. The sequence  $\{x_k\}$  is determined without ambiguity by  $x_1$  only; at iteration  $k$ , set

$$\mathbb{R}^k \ni \alpha \mapsto x(\alpha) := x_k + \sum_{i=1}^k \alpha_i s_i .$$

Thus, the method consists in minimizing with respect to  $\alpha \in \mathbb{R}^k$  the function  $q_k(\alpha) := f(x(\alpha))$ ; then, having a solution  $\bar{\alpha}$ , we set  $x_{k+1} = x(\bar{\alpha})$  and  $s_{k+1} = \nabla f(x(\bar{\alpha}))$ . Observe that

$$\frac{\partial q_k}{\partial \alpha_i}(\alpha) = \langle \nabla f(x(\alpha)), s_i \rangle \quad \text{for } i = 1, \dots, k$$

(and that all these derivatives must vanish for each  $k$  at  $\alpha = \bar{\alpha}$ ). It immediately follows that:

(i) for each  $k$ ,  $\langle s_{k+1}, s_i \rangle = 0$  for  $i = 1, \dots, k$  i.e.

$$\langle s_i, s_j \rangle = 0 \quad \text{for } i \neq j . \quad (2.4.2)$$

This implies:

(ii) the dimension of  $U_k$  increases by exactly one at each iteration (the gradients are independent!) until  $s_{k+1} = 0$  – an event which certainly happens for  $k \leq n$ , probably for  $k = n$ .

Setting  $x_{k+1} = x_k + t_k d_k$  as usual, one sees also that

- (iii)  $d_k$  has a nonzero component over  $s_k$  at each iteration (otherwise  $x_{k+1} = x_k$ , the minimum of  $f$  in  $V_{k-1}!$ ); as a result, the directions  $d_k$  are also independent,  $U_k$  can equivalently be defined as

$$U_k = \text{lin}\{d_1, \dots, d_k\} = \text{lin}\{d_1, \dots, d_{k-1}, s_k\}$$

and, for the same reason as in (i), there holds

$$\langle s_{k+1}, d_i \rangle = 0 \quad \text{for } i = 1, \dots, k. \quad (2.4.3)$$

- (iv) Because  $s(\cdot) = \nabla f$  is an affine mapping,  $s_{k+1} = s_k + t_k Q d_k$  holds and we obtain from (2.4.3)

$$0 = \langle s_k, d_i \rangle + t_k \langle Q d_k, d_i \rangle = t_k \langle Q d_k, d_i \rangle \quad \text{for } i = 1, \dots, k-1$$

which can be written

$$\langle s_{i+1} - s_i, d_k \rangle = 0 \quad \text{for } i = 1, \dots, k-1 \quad (2.4.4)$$

or also (remember (iii),  $t_k$  is nonzero):

$$\langle Q d_i, d_j \rangle = 0 \quad \text{for } i \neq j.$$

This last relation explains the name of the method: the directions are mutually *conjugate* with respect to the symmetric operator  $Q$ . Algebraically, we thus have a bi-orthogonalization process, which constructs the orthogonal sequence  $\{s_k\}$  and the sequence  $\{d_k\}$ , orthogonal with respect to the scalar product  $\langle\langle x, y \rangle\rangle := \langle x, Qy \rangle$ .

Computing  $d_k$  is an easy task: using (2.4.4), one writes

$$\langle s_i, d_k \rangle = \kappa_k \quad \text{for } i = 1, \dots, k \quad (2.4.5)$$

where the normalization factor  $\kappa_k$  is at our disposal ( $d_k$  is a direction!). Note that (2.4.5) is a linear system of  $k$  equations with  $k$  unknowns  $\alpha_i$  (see Definition 2.4.1). Thanks to (2.4.2), its solution is straightforward: we obtain  $d_k = \kappa_k \sum_{i=1}^k s_i / \|s_i\|^2$ .

It is useful to express  $d_k$  by recurrence formulae: writing

$$d_{k+1} = \frac{\kappa_{k+1}}{\kappa_k} d_k + \kappa_{k+1} \frac{s_{k+1}}{\|s_{k+1}\|^2}$$

and using the particular value  $\kappa_i = -\|s_i\|^2$ , we obtain

$$d_{k+1} = -s_{k+1} + \beta_k d_k \quad (2.4.6)$$

where

$$\beta_k = \frac{\|s_{k+1}\|^2}{\|s_k\|^2}. \quad (2.4.7)$$

With the above choice of the  $\kappa$ 's, the method becomes reminiscent of §2.2: the direction in (2.4.6) is that of the gradient plus a correction, which is a multiple of the previous direction.

The algorithm is now completely defined if we add that the stepsize must of course be optimal (a minimization in  $V_k$  of Definition 2.4.1 implies a minimization along  $d_k \in U_k$ !).

**Algorithm 2.4.2 (Conjugate Gradient, Linear Case)** The initial point  $x_1$  is given, as well as the tolerance  $\delta > 0$  and the objective function

$$f(x) = \frac{1}{2}\langle Qx, x \rangle + \langle b, x \rangle + c.$$

Again, the notation  $s := \nabla f$  is used. Set  $k = 1$ .

STEP 1 (stopping criterion). If  $\|s_k\| \leq \delta$  stop.

STEP 2 (computation of the direction). If  $k = 1$  set  $d_k = -s_k$ ; otherwise set  $d_k = -s_k + \beta_{k-1}d_{k-1}$ .

STEP 3 (line-search). Find  $t_k > 0$  solving

$$\min_{t>0} f(x_k + td_k) \quad \text{i.e.} \quad t_k = -\frac{\langle s_k, d_k \rangle}{\langle Qd_k, d_k \rangle}$$

and set  $x_{k+1} = x_k + t_k d_k$ .

STEP 4 ( $\beta$ -update and loop). Set  $\beta_k = \|s_{k+1}\|^2 / \|s_k\|^2$ , replace  $k$  by  $k + 1$  and loop to Step 1.  $\square$

**Remark 2.4.3** The tolerance  $\delta > 0$  for stopping may seem superfluous, since  $s_k$  must anyway be zero for some  $k \leq n + 1$ . However there are still two reasons for keeping  $\delta > 0$  as in the general case. First, the theoretical value  $s_k = 0$  may never be reached because of roundoff errors; second, if the number of variables is, say,  $n = 10^4$ , one is usually not prepared to wait for  $10^4$  iterations before stopping. In most applications,  $\|s_k\|$  is (fortunately) small enough long before that. Keeping in mind that conjugate-gradient methods are precisely tailored to large  $n$ , this remark is of course essential.  $\square$

**(b) Nonlinear Extensions** Now we should generalize the algorithm to non-quadratic objective functions. This is not straightforward because then, the theory breaks down from the very beginning. With regard to Definition 2.4.1, it is certainly not possible to minimize a general  $f$  (only known, recall, via the black box (U1) of Fig. 1.2.1) in any affine manifold – even one-dimensional, as is a direction.

A first simple idea is to pretend that  $f$  is quadratic and apply Algorithm 2.4.2 as it is – except that Step 3 must be set back to a general line-search scheme, to be seen in §3. At least one can reasonably hope that the direction of (2.4.6) is more efficient than the plain gradient because  $\beta_k > 0$  has a stabilizing effect. It smooths out the path linking the successive iterates (see Fig. 2.4.1: the angle between  $d_k$  and  $d_{k+1}$  is smaller than the angle between  $d_k$  and the gradient direction  $-s_{k+1}$ ).

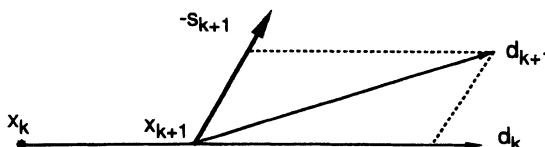


Fig. 2.4.1. Conjugate gradient as a path-smoothing device

Fortunately, there is a better argument. The value  $\beta_k$  of (2.4.7) comes out naturally from the system (2.4.5); there are several other possible values, however, which all

give the same result in the “perfect case” ( $f$  quadratic and optimal line-searches) but which are not equivalent in general. Let us mention

$$\hat{\beta}_k = \frac{\langle s_{k+1} - s_k, s_{k+1} \rangle}{\|s_k\|^2}, \quad (2.4.8)$$

known as the formula of Polak and Ribi  re (as opposed to (2.4.7), called the formula of Fletcher and Reeves). Observe the equivalence of (2.4.8) with (2.4.7), due to (2.4.2).

**Remark 2.4.4** It is amusing – and instructive – to mention the following point. The choice (2.4.7) is practically never used because (2.4.8) invariably converges faster. On the other hand, one can prove that, under suitable and reasonable assumptions on  $f$  and on the line-search, (2.4.7) yields a convergent algorithm, while a counter-example exists, showing that (2.4.8) need not converge to a critical point!  $\square$

Among the many possibilities to choose  $\beta_k$ , one is more interesting than the others, namely:

$$\tilde{\beta}_k = \frac{\langle s_{k+1} - s_k, s_{k+1} \rangle}{\langle s_{k+1} - s_k, d_k \rangle}.$$

To explain why, let us write the resulting direction in a rather artificial way. We set as in (2.3.8)  $\xi := x_{k+1} - x_k = t_k d_k$ ,  $\sigma := s_{k+1} - s_k$ . Now add a term which is 0 if the stepsize  $t_k$  is optimal (then  $\langle s_{k+1}, \xi \rangle = 0$ ):

$$\begin{aligned} d_{k+1} &= -s_{k+1} + \tilde{\beta}_k d_k = \\ &= -s_{k+1} + \frac{\langle \sigma, s_{k+1} \rangle}{\langle \sigma, d_k \rangle} d_k + \frac{\langle s_{k+1}, \xi \rangle}{\langle \sigma, \xi \rangle} \left[ \sigma - \left( 1 + \frac{|\sigma|^2}{\langle \sigma, \xi \rangle} \right) \xi \right]. \end{aligned} \quad (2.4.9)$$

We stress that, if the stepsizes are optimal, (2.4.9) provides another form of linear conjugate gradient, just as (2.4.6) – (2.4.7) or (2.4.8). The reason of our rather complicated form of (2.4.9) is the following. Do not suppose that  $f$  is quadratic nor that the stepsizes are optimal and consider the BFGS formula (2.3.10), with  $W$  replaced by the identity matrix. Then the resulting direction is exactly (2.4.9). In other words,  $\tilde{\beta}_k$  of (2.4.8) with the modification appearing in (2.4.9) (which matters only when  $t_k$  is not optimal) yields a “memoryless” quasi-Newton method in which, for want of storage, the sequence of matrices is reinitialized at each iteration.

This gives a serious motivation for the somewhat sophisticated formula (2.4.9), although it is slightly more expensive to compute than the others. Taken as a nonlinear conjugate-gradient method, it is based on a *second-order development* of  $f$ . By contrast, the purely algebraic arguments of the beginning of this Section 2.4 do not extend properly to the nonlinear case. Among its advantages, we mention that  $d_{k+1}$  of (2.4.9) is a descent direction under the mere condition  $\langle \sigma, \xi \rangle > 0$  (recall Theorem 2.3.6). This property is not easy to obtain with  $\beta$ -choices like (2.4.7) or (2.4.8).

**Remark 2.4.5** We mention that the way is thus open for methods intermediate between (2.3.9) and (2.4.9). Suppose one has a limited memory allowing the storage of say  $N$  real numbers simultaneously (with  $N \ll n^2/2$ ). Then a quasi-Newton method such as (2.3.10) cannot be used; but if we give up using the form (2.3.11) we see that, after all,  $W_k$  can be computed by using  $2k$  vectors only, say

$$\xi_1, \sigma_1, \xi_2, \sigma_2, \dots, \xi_k, \sigma_k.$$

Then, the intermediate methods we are alluding to would consist in using explicitly at each iteration the largest possible number  $K \simeq N/2n$  among the above pairs of vectors, so as to include in  $d_k$  as much information as possible concerning second-order behaviour of  $f$ ; observe in particular that (2.4.9) takes  $K = 1$ , while (2.3.10) demands  $K$  arbitrarily large.

**Remark 2.4.6** An easy observation will be of interest to us later in this book, when defining other generalizations of the conjugate-gradient method (§XIV.4.3). The direction obtained in the “perfect case” ( $f$  quadratic and optimal stepsize) is defined by (cf. (2.4.5))

$$d_k = \sum_{i=1}^k \bar{\alpha}_i s_i \quad \text{and} \quad \langle s_i, d_k \rangle = \langle s_j, d_k \rangle \quad \text{for } i, j = 1, \dots, k. \quad (2.4.10)$$

Geometrically, this means that  $d_k$  is orthogonal to the *affine* hull of the gradients, i.e. the affin hyperplane

$$H_k := \left\{ \sum_{i=1}^k \alpha_i s_i : \sum_{i=1}^k \alpha_i = 1 \right\}$$

of dimension  $k - 1$  passing through  $s_1, \dots, s_k$  (see Fig. 2.4.2). If  $d_k$  is scaled so that  $\sum_i \bar{\alpha}_i = -1$  in (2.4.10), then  $-d_k \in H_k$  and  $-d_k$  is actually the projection of the origin onto  $H_k$ .

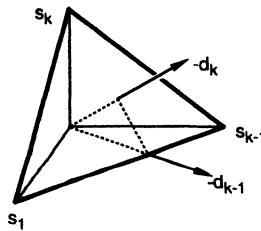


Fig. 2.4.2. Conjugation is projection

Furthermore, the orthogonality of the gradients also implies that the  $\bar{\alpha}_i$ 's defined by (2.4.10) all have the same sign and  $-d_k$  is actually the projection of the origin onto

$$C_k := \left\{ \sum_{i=1}^k \alpha_i s_i : \sum_{i=1}^k \alpha_i = 1, \alpha_i \geq 0 \right\},$$

the *convex* hull of the gradients  $s_1, \dots, s_k$ .

Note also the following property, again due to this orthogonality: if one (i) selects  $k - 1$  generators of  $H_k$ , say  $s_1, \dots, s_{k-1}$ , then (ii) projects the origin onto their convex hull  $C_{k-1}$  to obtain  $-d_{k-1}$ , and finally (iii) projects the origin onto the segment  $[-d_{k-1}, s_k]$ , one still obtains the same  $-d_k$  (see Fig. 2.4.2). This, among other things, explains why a recurrence formula like (2.4.6) is possible.  $\square$

### 3 Line-Searches

Now, considering the direction-finding problem as solved, we concentrate exclusively on the stepsize.

Thus, we are given a function defined by

$$\mathbb{R}^+ \ni t \mapsto q(t) := f(x_k + td_k),$$

where  $x_k$  is the starting point and  $d_k$  is the direction of search. More precisely – let us insist once again on this point – we are given a black box, namely (U1) in Fig. 1.2.1, which computes  $q(t)$  and  $q'(t) = \langle \nabla f(x_k + td_k), d_k \rangle$  for any value  $t \geq 0$  that we may choose (it is an elementary exercise to check that the derivative of  $q$  does have the above expression).

We also know that

$$q'(0) < 0$$

i.e.  $d_k$  is a descent direction in the sense of Definition 2.1.1.

With these data, we want to find a “suitable” (whatever this means) stepsize  $t > 0$ , satisfying in particular

$$q(t) < q(0). \quad (3.0.1)$$

Recalling Remark 1.2.3, we see that (3.0.1) implies a trial and error process, already mentioned in Remark 1.3.2. The line-search is in fact a subalgorithm, to be executed for each  $k$  at Step 3 of Algorithm 1.3.1. Our aim is now to study this subalgorithm, which is an essential element for an optimization program to work properly in practice. Actual computation of the direction is usually straightforward, even if its theory may require rather sophisticated mathematics but the situation is reversed here: only elementary mathematics is involved, and practical difficulties appear. It requires some computational expertise to implement a good line-search algorithm.

### 3.1 General Structure of a Line-Search

Just as any iterative algorithm, the line-search has a rule for iteration and a stopping criterion; this means that the following two questions must be answered:

- (i) How the sequence of trial stepsizes should it be computed?
- (ii) When is the current trial  $t$  acceptable as the real stepsize from  $x_k$  to  $x_{k+1}$ ?

**Remark 3.1.1** The line-search must also have an initialization: which  $t > 0$  should be tried first? Although it is extremely important, we pass it silently for the moment because it is really a question relevant to the direction-finding issue. Remark 3.4.2 will say a bit more about this problem.  $\square$

Between the above two items, the rule for iteration is the more important part for efficiency of an algorithm in general (after all, the stopping criterion is only executed once). For the line-search subalgorithm, however, it is the other way around. For one thing, (i) is a one-dimensional problem, relatively easy: on a line, there are only two alternatives, go right or go left. More importantly, the stopping rule for the line-search is (at least part of) *the iterative rule* for the outer descent algorithm, which directly conditions the choice of  $x_{k+1}$ . Furthermore, it is crucial that the line-search be stopped as soon as possible, since it is executed many times, as many times as there

are descent iterations. For these reasons, the concept of acceptable stepsize, i.e. the stopping criterion of the line-search, must be defined with great care.

Here, it is useful to organize the stopping criterion so that it gives not only a dichotomous answer by “yes” (the given stepsize is suitable) or “no” (it is not), but also some indication for the next trials in case the present one is not suitable. The *essential object* characterizing a line-search is then a *test* which, given a stepsize  $t$  and suitable information on  $q$  (obtained from (U1)), has three possible answers:

- (0) This  $t$  is suitable and the line-search can be stopped.
- (R) This  $t$  is not suitable and no suitable  $t^*$  is likely to be found on its right.
- (L) This  $t$  is not suitable and no suitable  $t^*$  is likely to be found on its left.

Then, any  $t$  for which (L) [resp. (R)] holds can serve as a lower [resp. upper] bound for subsequent trials because this  $t$  lies to the left [resp. to the right] of the interesting area of search. As a result, the following scheme suggests itself.

**Algorithm 3.1.2 (Schematic Line-Search)** An initial trial stepsize  $t > 0$  is given. Set  $t_L = 0$  and  $t_R = +\infty$ .

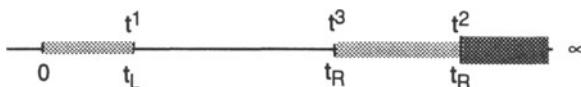
STEP 1. Test  $t$ ; in case (0) stop the line-search and pass to the next iterate  $x_{k+1}$ .

STEP 2. In case (L) set  $t_L = t$ ; in case (R) set  $t_R = t$ . Go to Step 3.

STEP 3. Select a new  $t$  in  $[t_L, t_R]$  and loop to Step 1. □

Some comments will help our understanding this procedure.

- During the early cycles, as long as  $t_R$  remains infinite, the update of Step 3 is actually an extrapolation beyond  $t_L$ ; when some real upper bound  $t_R < +\infty$  is found, the update becomes an interpolation between  $t_L$  and  $t_R$  (rather than the hard-to-define “ $t_R = +\infty$ ”, one could equally initialize “ $t_R = 0$ ”, meaning conventionally “no real upper bound has been found yet”).
- By construction,  $t_L$  can only increase,  $t_R$  can only decrease and all the  $t_L$ ’s generated during this process are strictly smaller than all the  $t_R$ ’s. The whole idea actually consists in generating a sequence of nested intervals  $[t_L, t_R]$ : at each cycle, either  $t_L$  or  $t_R$  is moved toward the other endpoint, thus reducing the interval. The interval  $[t_L, t_R]$  appears as a *safeguarding bracket*, inside which the final suitable  $t_k$  is searched for.



**Fig. 3.1.1. A possible line-search scenario**

Figure 3.1.1 illustrates a possible scenario: at the first trial  $t^1$ , suppose (L) holds; then  $t_L$  is moved from 0 to this  $t^1 > 0$ , which becomes a left-bound for all subsequent trials. The next trial is some  $t^2 > t^1$ ; suppose (R) holds there:  $t^2$  becomes a right-bound  $t_R$  and one takes  $t^3$  between  $t^1$  and  $t^2$ , i.e. between  $t_L$  and  $t_R$ . At  $t^3$ , it may be again (R), say, which holds, etc. Each trial increases the dashed area, in which no future trial is ever placed.

As was said earlier, it is *crucial* that Algorithm 3.1.2 be finite, i.e. that case (0) be reached after a finite, and if possible small, number of cycles. For this, two properties must be respected:

**Property 3.1.3 (Safeguard-Reduction Property)** The update of Step 3 must be such that:

- (i) infinitely many extrapolations would let  $t_L \rightarrow +\infty$ , and
- (ii) infinitely many interpolations would let  $t_R - t_L \rightarrow 0$ .  $\square$

**Property 3.1.4 (Consistency Property)** The test (O), (R), (L) must be organized so that:

- (i) when  $t$  is large enough, case (L) never occurs, and

- (ii) when  $t_R - t_L$  is small enough, case (0) occurs for all  $t$  in between.  $\square$

Properties 3.1.3(i) and 3.1.4(i) taken together imply that, after a finite number of cycles, either Algorithm 3.1.2 stops at case (0), or it finds a  $t_R < +\infty$ ; the number of extrapolations is finite. The (ii)-combination implies likewise that the number of interpolations is finite as well. Altogether, this implies that Algorithm 3.1.2 must terminate. For an easier understanding, another way to state 3.1.4(ii) is as follows: there is an interval  $I$  of positive length such that (0) occurs at any  $t \in I$ , and such that  $I \subset ]t_L, t_R[$  for all  $t_L$  and  $t_R$  generated by the algorithm.

**Example 3.1.5** The actual meaning of 3.1.4(ii) will become clearer later on; let us use a very simple and naive example to illustrate its importance. Define the test by:

- (0)  $t$  is suitable when  $q'(t) = 0$ ,
- (L)  $t$  is a  $t_L$  when  $q'(t) < 0$ ,
- (R)  $t$  is a  $t_R$  when  $q'(t) > 0$ .

This test is natural, as it is motivated by the desire to minimize  $q$  – i.e.  $f$  along  $d_k$ . The continuity of  $q'$  implies that, for arbitrary  $t_L$  and  $t_R$ , there is a  $t^*$  in between which satisfies (0) (the initialization  $t_L = 0$  is consistent thanks to the descent property  $q'(0) < 0$ ). Then Algorithm 3.1.2 should find such a  $t^*$ .

Yet, the resulting line-search is impractical, for at least two reasons:

- assuming for example that the equation  $q'(t) = 0$  has a unique solution, the interpolation process will never find it. Readers not familiar with numerical computations may not be convinced by this fact. They should think of the example  $q(t) = \exp t - 2t$ : no computer can find its minimum  $\log 2$ , an irrational number;
- even if, by an extraordinary luck, one lands on  $t$  satisfying (0), it may not be a minimum. It may not even satisfy  $q(t) < q(0)$ , see Fig. 3.1.2.  $\square$

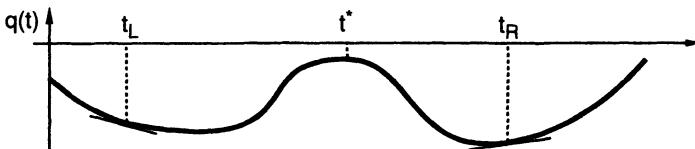


Fig. 3.1.2. An undesired stepsize

Since it is the more important, we will first study the test (O) (R) (L), which defines a suitable stepsize; then we will say a few words on the interpolation-extrapolation formulae for iterating the line-search subalgorithm.

### 3.2 Designing the Test (0), (R), (L)

Historically, the aim of the line-search has been to minimize the one-dimensional function  $q$ . Example 3.1.5, however, reveals that finding an optimal stepsize may take an infinite amount of computing time; approximating it is therefore likely to take a long time. Nowadays, it is recognized that, after all, since the function to be minimized is  $f$  and not  $q$ , it may not be a good idea to waste time on a minor, purely local problem. This is especially true of Newtonian methods, for which the stepsize  $t = 1$  is probably a better choice than an optimal stepsize – recall Remark 2.3.3.

**Remark 3.2.1** The total computing time of an iterative algorithm is roughly the average time spent by one iteration multiplied by the number of iterations. For a descent algorithm of the form 1.3.1, the number of iterations is mainly driven by the quality of the directions – and also by the stopping tolerance, say  $\delta$  of (1.2.1). As for the computing time of one iteration, it is the sum  $T_A + \ell T_U$ , where

- $T_A$  is the computing time needed by block (A) of Fig. 1.2.1,
- $\ell$  is the average number of trials needed by the line-search, and
- $T_U$  is the time needed for one execution of the black box (U1).

It is usually the case that  $T_U$  is much larger than  $T_A$ , say  $T_U \simeq 10T_A$ . The total execution time (of one iteration, hence of the overall descent algorithm) is therefore almost exclusively spent in (U1). Even more can be said: for practically all algorithms,  $T_A$  is small in absolute terms, say a fraction of a second on a “standard” computer. Thus, in those cases where  $T_U$  is not dominant,  $T_U$  is also small and trying to reduce the total execution time is not crucial: the net benefit will be again small, say a matter of seconds.

We conclude that the number of calls to (U1) is a sensible measure for the execution time of an optimization algorithm. It is therefore crucial for efficiency to keep  $\ell$  small (keeping in mind that the overall number of descent iterations – of line-searches – must also be kept small). Of course, the main ingredient influencing the value of  $\ell$  is the (O)-clause in the test. □

The modern point of view for designing a line-search consists in looking for a compromise between obtaining (0) fast, and decreasing  $f$  well. It can be added that, when something goes wrong in a descent algorithm of the form 1.3.1, it is invariably within the line-search, i.e. during the execution of Step 3. Another important aspect is therefore robustness: in addition to being *fast*, a line-search must be *fail-safe* and *simple*, so as to work as often as possible and, in case of failure, to make it clear where the possible troubles come from, as well as their possible cures.

In order to make the move from  $x_k$  to  $x_{k+1}$  reasonable, the least that can be required from a good stepsize is to be neither too large nor too small. A stepsize that is not too large is necessary to prevent the sequence  $\{x_k\}$  from oscillations, and in particular to force the decrease (3.0.1) (recall Fig. 1.3.1). On the other hand, the stepsize should not be too small, so as to yield a non-negligible progress from  $x_k$  toward the cluster point  $\bar{x}$ , which can be far if  $x_1$  is a poor initialization. Defining the test, and more precisely its (0)-part, consists precisely in giving a meaning to “large” and “small”.

**Deciding (R)** A fairly general consensus exists to define “large”: one says that  $t > 0$  is not too large when

$$q(t) \leq q(0) + mtq'(0) \quad (3.2.1)$$

where  $m$  is a coefficient chosen in  $]0, 1[$  (whose precise value is not crucial, various authors favour various choices, let us say for example that  $m = 0.1$  is a reasonable value).

In view of the descent requirement, property (3.2.1) makes much sense; in particular, it guarantees  $q(t) < q(0)$  automatically. Suppose that  $q$  were linear: then (3.2.1) would hold (even with  $m = 1$ ) for all  $t$ . As a result, (3.2.1) does hold when  $t > 0$  is close enough to 0, provided that  $q$  is smooth enough: the coefficient  $m < 1$  appears as a tolerance, allowing  $q$  to deviate from linearity.

A stepsize  $t$  satisfying “not (3.2.1)” is declared too large and case (R) occurs. Figure 3.2.1 illustrates all this: in dashed areas  $R_1$  and  $R_2$ , (R) occurs; they are both far from 0. For (0) to occur,  $t$  must be in a non-dashed area (although this is not sufficient: such a  $t$  may still be too small). To say that Property 3.1.4(i) does not hold is to say that the dashed area is bounded. This can happen only when  $q$  – hence  $f$  – is unbounded from below, implying that the minimization of  $f$  on  $\mathbb{R}^n$  is ill-posed and has no solution.

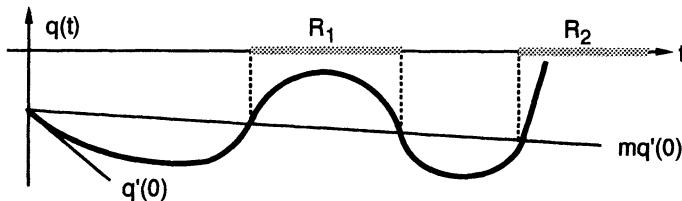


Fig. 3.2.1. Distributing the possible stepsizes

**Remark 3.2.2** The value  $m = 1/2$  in (3.2.1) plays a special role, due to a geometric property illustrated on Fig. 3.2.2: no matter where the point  $P$  is located on the parabola, the slope of its tangent is twice the slope of the chord pointing to the summit  $O$ :  $OM = OM'$ . In analytical terms, this property (which can easily be established) reads: if  $q$  is quadratic, and if  $t^*$  minimizes  $q$ , then

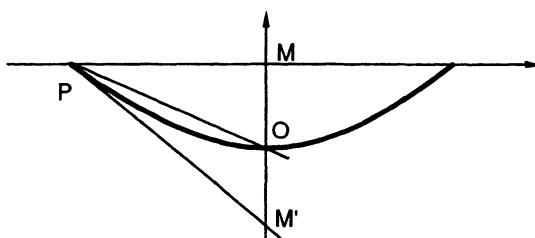


Fig. 3.2.2. A property of quadratic functions

$$q(t^*) = q(0) + \frac{1}{2}t^* q'(0).$$

This implies an important consequence: in (3.2.1), it is strongly advisable to take  $m < 1/2$ . Otherwise, when  $q$  happens to be quadratic, its optimal stepsize will be considered too large – a paradox.  $\square$

**Remark 3.2.3** Along the same lines, (3.2.1) has a useful interpretation. Let us come back to §2.3 – more precisely to (2.3.1) and (2.3.2) – and let us interpret the function

$$\tilde{f}(d) := f(x_k) + \langle s(x_k), d \rangle + \frac{1}{2}\langle Qd, d \rangle$$

as a *model*, which approximates  $f(x_k + d)$ . The strategy of §2.3 consists of minimizing this model to obtain  $d_k$ , the direction of search. Then we obtain the restricted model

$$\mathbb{R} \ni t \mapsto \tilde{q}(t) := \tilde{f}(x_k + td_k) = q(0) + tq'(0) + \frac{1}{2}at^2$$

where  $a := \langle d_k, Qd_k \rangle$ . With this interpretation, the idea of the line-search is to correct the model  $\tilde{f}$  by a further adjustment of  $t > 0$ . Of course,  $d_k := -Q^{-1}s_k$  is such that  $\tilde{q}$  is minimized at  $t = 1$  (the minimum of  $\tilde{q}$  has to reproduce the minimum of  $\tilde{f}$ !). In fact, direct calculations show that  $\tilde{q}$  can be written

$$\tilde{q}(t) = q(0) + tq'(0) - \frac{1}{2}t^2q''(0). \quad (3.2.2)$$

Now suppose we are testing  $t = 1$  in (3.2.1); we ask the question:

$$\text{is } q(1) - q(0) \text{ lower than } mq'(0) ?$$

which, using the form (3.2.2) for  $\tilde{q}$ , can be written:

$$\text{is } q(1) - q(0) \text{ lower than } 2m[\tilde{q}(1) - q(0)] ?$$

In other words, the question that we are asking is: “Does the model  $\tilde{f}$  agree well enough with the real  $f$ ?”; or also: “Is the real decrease  $q(1) - q(0)$  at least a fraction (namely  $2m$ , supposedly smaller than 1) of the predicted decrease  $\tilde{q}(1) - q(0)$ ?

With this interpretation,  $q'(0)$  plays a minor role and is rather replaced by  $\tilde{q}(1) - q(0)$ . This will be useful later on (Sections XV.1.3 and XV.3.3(c)) in special situations when  $q'(0)$  is actually unknown. Of course, the idea could be extended to stepsizes other than 1: for example, one could imagine replacing the descent test (3.2.1) by

$$q(t) \leq q(0) + 2m[\tilde{q}(t) - q(0)].$$

This cannot be done without care, however: if  $\tilde{q}(t) \geq q(0)$ , the descent property  $q(t) < q(0)$  is no longer guaranteed.  $\square$

**Deciding (L)** It remains to define the stepsizes that are too small. Here there are several possibilities and we just mention two of them, namely (3.2.3) and (3.2.4) below. One chooses another tolerance  $m'$  satisfying

$$0 < m < m' < 1;$$

then, for some authors,  $t$  is not too small when

$$q(t) \geq q(0) + m'tq'(0) \quad (3.2.3)$$

while for some others,  $t$  is not too small when

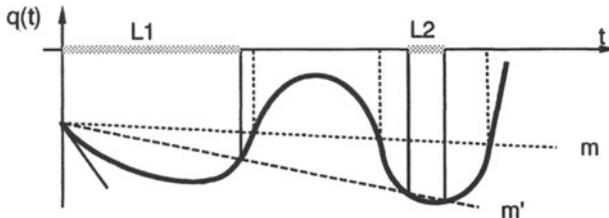
$$q'(t) \geq m'q'(0). \quad (3.2.4)$$

Inspect Fig. 3.2.1 carefully for an interpretation. Neither (3.2.3) nor (3.2.4) can hold for  $t > 0$  close to 0; this is due to the continuity of  $q'$  and to the descent property  $q'(0) < 0$ .

To sum up, there are (at least) two possibilities to design (0), (R), (L). One is

- (0) holds when  $t$  satisfies (3.2.1) and (3.2.3),
- (R) holds when  $t$  does not satisfy (3.2.1),
- (L) holds when  $t$  does not satisfy (3.2.3),

sometimes called the criterion of Goldstein and Price and illustrated in Fig. 3.2.3. With relation to 3.1.4(ii), this figure shows rather clearly that, for two arbitrary points  $t_L$  and  $t_R$  with  $t_L < t_R$ , there is always an (0)-segment in between.



**Fig. 3.2.3.** The criterion of Goldstein and Price

The other possibility is

$$\left. \begin{array}{l} (0) \text{ holds when } t \text{ satisfies (3.2.1) and (3.2.4),} \\ (R) \text{ holds when } t \text{ does not satisfy (3.2.1),} \\ (L) \text{ holds when } t \text{ satisfies (3.2.1) but not (3.2.4),} \end{array} \right\} \quad (3.2.5)$$

which is often called the Wolfe criterion. We observe that, when  $q(t)$  is very small,  $t$  seems excellent for a descent to  $f(x_{k+1})$ ; but (3.2.3) does not hold, so case (L) occurs and this  $t$  is rejected by the criterion of Goldstein and Price. This appears as a deficiency and it explains that Wolfe's criterion is usually preferred. On the other hand, the criterion of Goldstein and Price does not require computing  $q'$ . It is therefore useful when  $f$  can be computed much more cheaply than  $\nabla f$  in the black box (U1) (remember Remark 3.2.1); this situation may happen in practice, although rarely.

### 3.3 The Wolfe Line-Search

For an illustration we give the specific form of Algorithm 3.1.2 when the test (3.2.5) is used. It is not only useful for the descent methods outlined in this chapter, but it is also convenient for more sophisticated methods, to be seen later in this book. Therefore, it deserves special study.

**Algorithm 3.3.1 (Wolfe's Line-Search)** An initial trial  $t > 0$  is given, as well as  $m \in ]0, 1[$  and  $m' \in ]m, 1[$ . Set  $t_L = 0$  and  $t_R = 0$ .

STEP 1 (Test for large  $t$ ). Compute  $q(t)$ ; if (3.2.1) does not hold set  $t_R = t$  and go to Step 4.

STEP 2 (Stopping criterion;  $t$  is not too large). Compute  $q'(t)$ ; if (3.2.4) holds stop the line-search and pass to the next iterate  $x_{k+1}$ . Otherwise set  $t_l = t$  and go to Step 3.

STEP 3 (Extrapolation). If  $t_R > 0$  go to Step 4.

Otherwise find a new  $t$  by extrapolation beyond  $t_L$  and loop to Step 1.

STEP 4 (Interpolation). Find a new  $t$  by interpolation in  $]t_L, t_R[$  and loop to Step 1.  $\square$

Given :  $q(0), q'(0) < 0, 0 < m < m' < 1, t > 0$ .

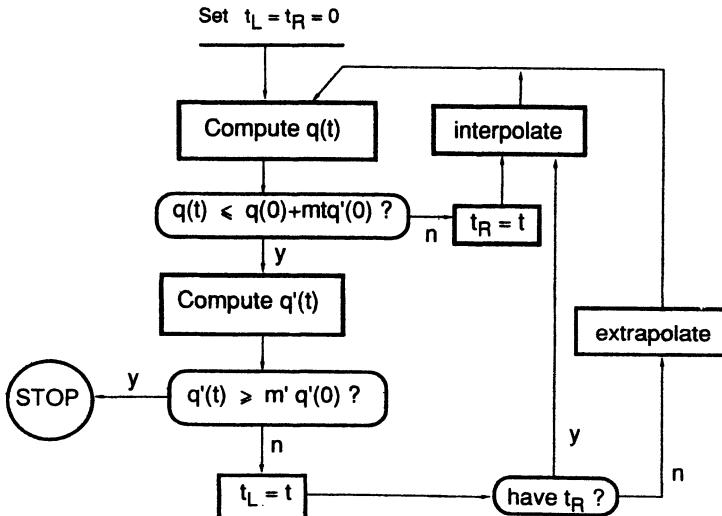


Fig. 3.3.1. Wolfe's line-search

The flow-chart corresponding to this algorithm is displayed in Fig. 3.3.1. It is fairly simple, which means that corresponding computer programs can be made fairly robust. We show now that the consistency property 3.1.4 holds:

**Theorem 3.3.2 (Consistency of Wolfe's Line-Search)** Assume that  $q$  is continuously differentiable, and that the descent property  $q'(0) < 0$  and the safeguard-reduction property 3.1.3 hold. Then, Algorithm 3.3.1 either generates a sequence of stepsizes  $t$  with  $q(t) \downarrow -\infty$ , or terminates after a finite number of cycles.

PROOF. Suppose that the stop never occurs. First observe that we have at each cycle

$$q(t_L) \leq q(0) + mt_L q'(0). \quad (3.3.1)$$

Suppose first that Algorithm 3.3.1 loops indefinitely between Step 3 and Step 1. Then, by construction, every generated  $t$  is a  $t_L$  and satisfies (3.3.1). From 3.1.3(i),  $t_L \rightarrow +\infty$ ; because  $q'(0) < 0$ , (3.3.1) shows that  $q(t_L) \rightarrow -\infty$ .

Thus, if  $f(x + td)$  is bounded from below,  $t_R$  becomes positive at some cycle. From then on, the algorithm loops between Step 4 and Step 1; by construction, we have at each subsequent cycle

$$q(t_R) > q(0) + mt_R q'(0), \quad (3.3.2)$$

the sequence  $\{t_L\}$  is (strictly) increasing, the sequence  $\{t_R\}$  is (strictly) decreasing, every  $t_L$  is smaller than every  $t_R$ , and Property 3.1.3(ii) implies that these two sequences are actually adjacent i.e., for some  $t^* \geq 0$ :

$$t_L \uparrow t^* \quad \text{and} \quad t_R \downarrow t^*;$$

(3.3.1) and (3.3.2) imply by the continuity of  $q$  that

$$q(t^*) = q(0) + mt^* q'(0). \quad (3.3.3)$$

Then we write (3.3.2) as

$$q(t_R) > q(0) + mq'(0)(t_R - t^* + t^*) = q(t^*) + m(t_R - t^*)q'(0).$$

By (3.3.2) and (3.3.3),  $t_R > t^*$ , hence

$$\frac{q(t_R) - q(t^*)}{t_R - t^*} > mq'(0)$$

and, passing to the limit:

$$q'(t^*) \geq mq'(0) > m'q'(0),$$

where we have used  $q'(0) < 0$  and  $m < m'$ .

Now, the stopping criterion of Step 2 implies that  $q'(t_L) < m'q'(0)$  and it suffices to pass to the limit to obtain the contradiction

$$q'(t^*) \geq mq'(0) > m'q'(0) \geq q'(t^*). \quad (3.3.4)$$

□

We leave it as an exercise to show the same result for the criterion of Goldstein and Price, simply by modifying the end of the above proof.

**Remark 3.3.3** It is only a rather weak continuity property of  $q'$  that is used in the above proof. It serves only to obtain the contradiction (3.3.4), and for this a *left continuity* property alone is needed for  $q'$  (note that  $t_L \uparrow t^*$ !). This remark, which we leave informal for the moment, will become essential for the numerical methods considered in this book. □

**Remark 3.3.4** With relation to the secant methods of §2.3, we mention an important property of the Wolfe criterion. In view of (3.2.4), the actual stepsize  $t_k$  satisfies  $q'(t_k) \geq m'q'(0)$ , which can be written

$$q'(t_k) - q'(0) \geq (m' - 1)q'(0) > 0$$

or, using the gradients explicitly:

$$\langle s(x_{k+1}) - s(x_k), d_k \rangle \geq (m' - 1)\langle s(x_k), d_k \rangle > 0. \quad (3.3.5)$$

Now recall Theorem 2.3.6: the quasi-Newton operators  $W_k$  in Algorithm 2.3.4 remain positive definite if (and only if)  $\langle \sigma_k, \xi_k \rangle > 0$  for all  $k$  (the notation (2.3.8) is used); knowing that  $\xi_k = t_k d_k$ , this is clearly implied by (3.3.5), which can be written

$$\frac{1}{t_k} \langle \sigma_k, \xi_k \rangle \geq \frac{m'-1}{t_k} \langle s(x_k), \xi_k \rangle > 0.$$

In other words, the Wolfe criterion automatically preserves positive definiteness in secant methods, hence its interest in this framework.  $\square$

It remains to check that our general methodology for line-searches – and in particular the Wolfe criterion (3.2.1), (3.2.4) – does preserve the convergence properties of the outer descent algorithm. This question cannot be answered in absolute terms since it also depends on the choice of the direction. Convergence results have to be stated for each combination “direction  $\times$  stepsize”. To illustrate how they are proved, we consider the following realization of Algorithm 1.3.1:

**Algorithm 3.3.5 (Steepest-Descent with Wolfe’s Line-Search)** The starting point  $x_1 \in \mathbb{R}^n$  and the tolerances  $\delta > 0$ ,  $0 < m < m' < 1$  are given, as well as the line-search subalgorithm 3.3.1. Set  $k = 1$ .

STEP 1. If  $\|\nabla f(x_k)\| \leq \delta$  stop.

STEP 2. Take  $d_k = -\nabla f(x_k)$ .

STEP 3. Obtain  $x_{k+1} = x_k + t_k d_k$  satisfying

$$f(x_{k+1}) \leq f(x_k) - mt_k \|\nabla f(x_k)\|^2 \quad (3.3.6)$$

$$\langle \nabla f(x_{k+1}), d_k \rangle \geq -m' \|\nabla f(x_k)\|^2. \quad (3.3.7)$$

STEP 4. Replace  $k$  by  $k + 1$  and loop to Step 1.  $\square$

Of course, this algorithm is given for the sake of illustration only, since the gradient method is *forbidden* (recall the end of §2.2). In fact, Algorithm 3.3.5 is good enough for our present purpose, which is to demonstrate the convergence mechanism of a descent method.

**Theorem 3.3.6** Suppose that  $\nabla f$  is uniformly continuous on the sublevel-set

$$S_{f(x_1)}(f) := \{x \in \mathbb{R}^n : f(x) \leq f(x_1)\}.$$

Then Algorithm 3.3.5 stops in Step 1 for some finite  $k$ , unless  $f(x_k) \rightarrow -\infty$ .

PROOF. We proceed in three steps.

[*(i)*] From (3.3.6), we get

$$f(x_k) - f(x_{k+1}) \geq m \|\nabla f(x_k)\| \|t_k d_k\| = m \|\nabla f(x_k)\| \|x_{k+1} - x_k\|. \quad (3.3.8)$$

[*(ii)*] On the other hand, subtracting  $\langle \nabla f(x_k), d_k \rangle = -\|\nabla f(x_k)\| \|d_k\|$  from (3.3.7), we get

$$\langle \nabla f(x_{k+1}) - \nabla f(x_k), d_k \rangle \geq (1 - m') \|\nabla f(x_k)\| \|d_k\|.$$

So, using the Cauchy-Schwarz inequality:

$$\|\nabla f(x_{k+1}) - \nabla f(x_k)\| \geq (1 - m') \|\nabla f(x_k)\|. \quad (3.3.9)$$

[*(iii)*] Now assume that the algorithm does not stop ( $\|\nabla f(x_k)\| > \delta$  for all  $k$ ) and that  $\{f(x_k)\}$  is bounded from below. Then both sides in (3.3.8) form a convergent series;  $x_{k+1} - x_k \rightarrow 0$ ; the uniform continuity of  $\nabla f$  implies  $\nabla f(x_{k+1}) - \nabla f(x_k) \rightarrow 0$ , a contradiction to (3.3.9).  $\square$

Observe the scheme of this proof:  $\|\nabla f(x_k)\|$  plays the role of a convergence parameter, hopefully tending to 0; then (i) [resp. (ii)] quantifies the fact that  $t_k$  is not too large [resp. not too small]; finally (iii) synthesizes these arguments. If the stopping criterion of Step 1 were not present, the same scheme would prove by contradiction that 0 is a cluster point of the sequence  $\{\nabla f(x_k)\}$  – remember (1.1.8).

### 3.4 Updating the Trial Stepsize

We turn now to the possibilities for interpolation and extrapolation in an algorithm such as 3.3.1. The simplest way to satisfy the safeguard-reduction property 3.1.3 is a rough doubling and halving process such as

- if  $t_R = +\infty$  replace  $t$  by  $2t$ ,
- if  $t_R < +\infty$  replace  $t$  by  $1/2(t_L + t_R)$ .

More intelligent formulae exist, however: as the number of cycles increases in the line-search, more and more information is accumulated from  $q$ , which can be used to guess where a convenient  $t$  is likely to lie. Then, the idea is to fit a simple model-function (like a polynomial) to this information. The model-function is used to obtain a *desired value*, say  $t_d$ , for the next trial, and it remains to force  $t_d$  inside the safeguard  $[t_R, t_L]$ , so as to ensure the safeguard-reduction property 3.1.3.

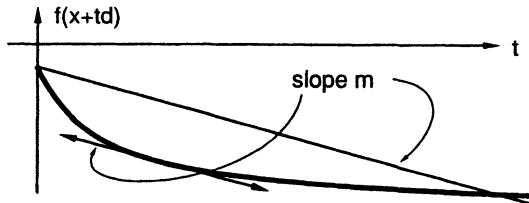
**Remark 3.4.1** The idea of having  $q(t) := f(x_k + td_k)$  as line-search function, of fitting a model to it, say  $\theta(t)$ , and of choosing  $t_d$  minimizing  $\theta$ , is attractive but may not be the most suitable. Remember that the descent test (3.2.1) might be satisfied by no minimizer of  $q$ .

A possible way round this discrepancy is to compute  $t_d$  minimizing the tilted function  $t \mapsto \theta(t) - mtq'(0)$ ; or equivalently to choose  $\theta(t)$  fitting  $q(t) - mtq'(0)$ ; or also to take  $q(t) := f(x_k + td_k) - mt\langle \nabla f(x_k), d_k \rangle$  as line-search function. The resulting  $t_d$  will certainly aim at satisfying (3.2.1) and

$$\langle \nabla f(x_k + td_k), d_k \rangle \geq m \langle \nabla f(x_k), d_k \rangle.$$

It will thus aim at satisfying (3.2.4) as well, and this strategy is more consistent with a Wolfe criterion, say; see Fig. 3.4.1.

The above strategy may look anti-natural, but luckily the perturbation term  $mtq'(0)$  is small, admitting that  $m$  itself is small (see Remark 3.2.2).  $\square$



**Fig. 3.4.1.** A perturbation of the line-search function

**(a) Forcing the Safeguard-Reduction Property** The forcing mechanism can be done as follows:

- When no  $t_R > 0$  has been found yet, one chooses  $\kappa > 1$  and the next trial  $t_+$  is  $\max\{t_d, \kappa t_L\}$  (each extrapolation multiplies the stepsize at least by  $\kappa$ ;  $\kappa$  may vary with the number of extrapolations but one should not let it tend to 1 without precautions).
- When some  $t_R < +\infty$  is on hand, one chooses  $\rho \in ]0, 1/2]$  and one does the following:
  - replace  $t_d$  by  $\max\{t_d, (1 - \rho)t_L + \rho t_R\}$ ;
  - then replace the new  $t_d$  thus obtained by  $\min\{t_d, \rho t_L + (1 - \rho)t_R\}$ ;
  - finally, the next trial  $t_+$  is set to this last  $t_d$ .

In other words,  $t_+$  is forced inside the interval obtained from  $[t_L, t_R]$  by chopping off  $\rho(t_R - t_L)$  from its two endpoints. At the next cycle,  $t_+$  will become a  $t_L$  or a  $t_R$  (unless clause (0) occurs) and in both cases, the bracket  $[t_L, t_R]$  will be reduced by a factor of at least  $1 - \rho$ ;  $\rho$  may vary at each interpolation but one must not let  $\rho \downarrow 0$  without precaution.

These questions, particularly that of choosing  $t_d$ , present a moderate interest because efficient line-searches need on the average far less than two cycles to reach (0) and to accomplish the descent iteration. Asymptotic properties of the interpolation formulae are therefore hardly involved.

**Remark 3.4.2** The question of the initial trial is crucial, since the above-mentioned score of less than two cycles per line-search is certainly not attainable without a good initialization. With Newtonian methods, one *must* try  $t = 1$  first, so the Newton step has a chance to prevail. For other methods, the following technique is universally used: pretend that  $q$  is quadratic

$$q(t) \simeq \frac{1}{2} a t^2 + q'(0)t + q(0) \quad (a > 0 \text{ is unknown})$$

and that its decrease from  $t = 0$  to the (asserted) optimal  $t^* := -q'(0)/a$  is going to be  $\Delta := f(x_{k-1}) - f(x_k)$ ; it is straightforward to check that  $t^*$  is then given by  $t^* = -2\Delta/q'(0)$ , which happens to be an excellent initialization.

Observe that, at the first descent iteration  $k = 1$ ,  $\Delta$  does not exist. In the notations of Fig. 1.2.1, it is actually block (U0) which should give to block (A) an idea of the very first initial trial; for example (U0) may pass to (A) an estimate of  $\Delta$  together with  $x_1$  and  $\delta$ . These are the kind of details that help an optimization program to run efficiently.  $\square$

**Remark 3.4.3** Having thus settled the question of the initialization, let us come again to the question of stopping criteria. The “ideal” event  $\|\nabla f(x_k)\| \leq \delta$  occurs rarely in practice, for many possible reasons. One is that  $\delta$  may have been chosen too small by the user and, in view of the speed of the minimization algorithm, iterations should go on essentially forever.

Another reason, a very common one, is that  $s(x)$  is actually not the gradient of  $f$  at  $x$ , either (i) because of a mistake in the black box (U1) – this is fairly frequent, see Remark 1.2.2 – or (ii) simply because of roundoff errors: (U1) can work only with finitely many digits, and there must be a threshold under which the computation errors become important. Then, observed values of  $q$  and of  $q' [= \langle s, d \rangle]$  are inconsistent with each other and a proof like that of Theorem 3.3.2 does not reflect reality. For example, the property

$$\frac{q(t_R) - q(t^*)}{t_R - t^*} \rightarrow q'(t^*) \quad \text{when } t_R \downarrow t^*$$

may become totally wrong. As a result, (0) never occurs,  $t_R - t_L$  does tend to zero and the line-search loops forever. The cause of this problem can be (i) or (ii) above; in both cases, the process must be stopped manually, so as to “spare” computing time (i.e. reduce it from infinity to a reasonable value!).

Thus, in addition to  $\delta$  for the ideal test (1.2.1), the user [i.e. block (U0)] must set another tolerance, say  $\delta'$ , allowing block (A) to somehow guess what a very small stepsize is. This  $\delta'$  defines a threshold, under which  $t_R - t_L$  must be considered as essentially 0. In these conditions, an “emergency stopping criterion”, acting when  $t_R - t_L$  becomes lower than this threshold, can be inserted in Step 4 of Algorithm 3.3.1. Note also that another emergency stop can be inserted in Step 3 to prevent infinite loops occurring when the objective function is unbounded from below.

It is interesting to observe that this  $\delta'$ -precaution is not sufficient, however: the user may have overestimated the accuracy of the calculations in (U1) and  $\delta'$  may never act, again because of roundoff errors. There exists, at last, an unfailing means for Algorithm 3.3.1 to detect that it is starting to loop forever. When the new stepsize  $t_+$  becomes close enough to a previous one, say  $t_L$ , there holds

$$x_k + t_L d_k = x_k + t_+ d_k$$

although  $t_L \neq t_+$ . This is another effect of roundoff errors – but beneficial, this time: when it happens, Algorithm 3.3.1 can be safely stopped.

All this belongs more to the art of computer programming than to mathematics and explains what we meant in the introduction of this Section 3, when mentioning that implementing a good line-search requires experience.  $\square$

We gave the above details in Remark 3.4.3 because they illustrate the kind of care that must be exercised when organizing automatic calculations. We conclude this section with some more details concerning the fit of  $q$  by some simple function. Although not particularly exciting, they are further illustrations of another kind of precaution: when doing a calculation, one should try to avoid division by 0!

**(b) Computing the Interpolation  $t_d$**  The most widely used fit for  $q$  is by a cubic function, which is done by the following calculations:

- Call  $\alpha$  and  $\alpha_-$  the two stepsize-values that have been tried last (the current one and the previous one; at the first cycle,  $\alpha_- = 0$ ).
- We have on hand  $q := q(\alpha)$ ,  $q' := q'(\alpha)$ ,  $q_- := q(\alpha_-)$  and  $q'_- := q'(\alpha_-)$ .

- These four data define a polynomial of degree  $\leq 3$  in  $t$ , which we find convenient to write as

$$\theta(t) := \frac{1}{3}a(t - \alpha)^3 + b(t - \alpha)^2 + q'(t - \alpha) + q.$$

- The coefficients  $a$  and  $b$  are identified by equating  $\theta(\alpha_-)$  and  $\theta'(\alpha_-)$  with  $q_-$  and  $q'_-$  respectively. Knowing that  $E := \alpha - \alpha_- \neq 0$  this gives the linear system

$$\begin{aligned}\frac{1}{3}E^2a - Eb &= Q' - q' \\ E^2a - 2Eb &= q'_- - q'\end{aligned}$$

in which we have set  $Q' := (q - q_-)/E$ . With  $P' := q' + q'_- - 3Q'$ , its unique solution is

$$E^2a = q' + q'_- + 2P' \quad \text{and} \quad Eb = q' + P'.$$

- Then the idea is to take  $t_d$  as the local minimum of  $\theta$  (if it exists), i.e. one of the real solutions (if they exist) of the equation

$$\theta'(t) = a(t - \alpha)^2 + 2b(t - \alpha) + q' = 0. \quad (3.4.1)$$

With respect to the unknown  $t - \alpha$ , the reduced discriminant of this equation is

$$\Delta := b^2 - aq' = \frac{1}{E^2}(P'^2 - q'q'_-) \quad (3.4.2)$$

which we assume nonnegative, otherwise there is nothing to compute.

- Clearly enough, if  $t - \alpha = (-b \pm \Delta^{1/2})/a$  solves (3.4.1), then

$$\theta''(t) = 2a(t - \alpha) + 2b = \pm 2\Delta^{1/2}.$$

Because  $\theta''$  must be nonnegative at  $t_d$ , it is the “+” sign that prevails; in a word,  $t_d$  can be computed by either of the following equivalent formulae:

$$t_d - \alpha = \frac{-b + \Delta^{1/2}}{a} \quad (3.4.3)$$

$$t_d - \alpha = \frac{(-b + \Delta^{1/2})(b + \Delta^{1/2})}{a(b + \Delta^{1/2})} = \frac{-q'}{b + \Delta^{1/2}}. \quad (3.4.4)$$

- The tradition is to use (3.4.3) if  $b \leq 0$  and (3.4.4) if  $b > 0$ . Then, roundoff errors are reduced because the additions involve two nonnegative numbers. In particular, the denominator in (3.4.4) cannot be zero.
- Now comes the delicate part of the calculation. In both cases, the desired  $t$  is expressed as

$$t_d = \alpha + \frac{N}{D} \quad (3.4.5)$$

but this division may blow up if  $D$  is close to 0. On the other hand, we know that if  $t_d$  is going to be outside the interval  $]t_L, t_R[$  (assumed to be known; in case of extrapolation we can temporarily set  $t_R = 10\kappa t_L$ ), the forcing mechanism of §(a) above will kill the computation of  $t_d$ . A formula like (3.4.5) is then useless anyway.

Now, a key observation is that  $\alpha \in [t_L, t_R]$ . In fact, the current trial  $\alpha$  is either  $t_L$  or  $t_R$ , as can be seen from a long enough contemplation of Algorithm 3.1.2. Then the property

$$t_L - \alpha < \frac{N}{D} < t_R - \alpha,$$

which must be satisfied by  $t_d$ , implies

$$\frac{|N|}{|D|} < t_R - t_L. \quad (3.4.6)$$

To sum up,  $t_d$  should be computed from (3.4.5) only if (3.4.6) holds. Otherwise the cubic model is helpless, as it predicts a new stepsize outside the bracket  $[t_L, t_R]$ . Then  $t_d$  should be set for example to  $t_L$  or  $t_R$  according to the sign of  $q'$ .

**Remark 3.4.4** Perhaps the most important reason for these precautions is the danger of so-called overflows. When a computer is asked to do an arithmetic operation whose result is very large in absolute value – say larger than  $10^{50}$  – it usually stops. All the current work is definitely lost, with all its intermediate results; a disaster (note here that the situation is not symmetric: if the result is small in absolute value, then it can be replaced by 0; in other words, a computer understands 0 but not “infinity”). This danger appears when multiplications, or equivalently divisions, are done; additions are less critical.

In our context, there are two types of quantities:  $t$ -values and  $q$ -values; they are independent in the sense that they are expressed in different units; their ratios form a third type:  $q'$ -values. One may think, for example, that  $t \simeq 10^{-10}$ ,  $q \simeq 10^{20}$ , so  $q' \simeq 10^{30}$ . Then, the above calculations should be performed with some care.

Because  $Q'$  is homogeneous to a derivative, its computation is relatively safe. By contrast, (3.4.2) is dangerous because terms like  $P'^2$  may have crazy values: in our example above,  $P'^2 \simeq 10^{60}$ . Thus, computing directly  $P'^2 - q'q'_-$  should be done only if  $|P'| \leq 1$ . Otherwise, observing that only  $\Delta^{1/2}$  is used, one should write (3.4.2) as (we skip the sign-problems)

$$\sqrt{P'^2 - q'q'_-} = \sqrt{P'} \sqrt{P' - (q'/P')q'_-}$$

and respect the stated order when computing this right-hand side. Finally, the test (3.4.6) is necessary only when  $|D| \leq 1$ ; then, it should be performed as

$$|N| \leq |D| (t_R - t_L).$$

□

**Remark 3.4.5** The distinction between (3.4.3) and (3.4.4) reduces the roundoff errors, and it also takes care of a vanishing  $a$ . The event  $a = 0$  does happen from time to time, namely when  $q$  is (close to) a quadratic function. From this point of view, the role of the sign of  $b$  is essential.

– If  $b > 0$ , then formula (3.4.4) is used and the role of  $a$  is minor: even if  $a \simeq 0$  (meaning that  $q$  looks like a convex quadratic function) (3.4.4) gives the safe value  $t_d \simeq -q'/b$ .

- If  $b \leq 0$ , then the case  $a \simeq 0$  means that  $\theta$  looks like concave quadratic, possibly linear ( $b = 0$ ); in both cases,  $\theta$  has no minimum and it is the cubic approximation which is meaningless anyway.

The interesting point is that this whole technique gives a stable computation of  $t_d$ , even though the cubic function  $\theta$  may degenerate. It should be said that the computation (3.4.2) of  $\Delta$  does suffer from roundoff, though:  $Q' = q'(\tau)$  for some  $\tau \in ]\alpha, \alpha_-[$ , so  $P'$  and  $\Delta$  are obtained by subtracting numbers which may have the same sign and be close together. When  $\alpha$  and  $\alpha_-$  are close together,  $\Delta$  has to be close to 0 even though  $q'$ -values are large.  $\square$

### III. Convex Sets

**Prerequisites.** Topology and Euclidean geometry of  $\mathbb{R}^n$ ; a skill to visualize 2- and 3-dimensional objects, so as to illustrate the results, and to support the intuition in higher dimensions.

**Introduction.** Our working space is  $\mathbb{R}^n$ . We recall that this space has the structure of a real vector space (its elements being called *vectors*), and also of an affine space (a set of *points*); the latter can be identified with the vector-space  $\mathbb{R}^n$  whenever an *origin* is specified. It is not always possible, nor even desirable, to distinguish vectors and points.

We equip  $\mathbb{R}^n$  with a scalar product  $\langle \cdot, \cdot \rangle$ , so that it becomes a Euclidean space, and also a complete normed vector space for the norm  $\|x\| := \sqrt{\langle x, x \rangle}$ . If an orthonormal basis is chosen, there is no loss of generality in assuming that  $\langle x, y \rangle$  is the usual dot-product  $x^\top y$ ; see §A.3.

The concepts presented in this chapter are of course fundamental, as practically all the subsequent material is based on them (including the study of convex functions). These concepts must therefore be fully mastered, and we will insist particularly on ideas, rather than technicalities.

## 1 Generalities

### 1.1 Definition and First Examples

**Definition 1.1.1** The set  $C \subset \mathbb{R}^n$  is said to be convex if  $\alpha x + (1 - \alpha)x'$  is in  $C$  whenever  $x$  and  $x'$  are in  $C$ , and  $\alpha \in ]0, 1[$  (or equivalently  $\alpha \in [0, 1]$ ).  $\square$

Geometrically, this means that the line-segment

$$[x, x'] := \{\alpha x + (1 - \alpha)x' : 0 \leq \alpha \leq 1\}$$

is entirely contained in  $C$  whenever its endpoints  $x$  and  $x'$  are in  $C$ . Said otherwise: the set  $C - \{c\}$  is a *star-shaped* set whenever  $c \in C$  (a star-shaped set is a set containing the segment  $[0, x]$  for all its points  $x$ ). A consequence of the definition is that  $C$  is also path-connected, i.e. two arbitrary points in  $C$  can be linked by a continuous path.

**Examples 1.1.2 (Sets Based on Affinity)** We have seen in Chap. I that the convex sets in  $\mathbb{R}$  are exactly the intervals; let us give some more fundamental examples in several dimensions.

- (a) An *affine hyperplane*, or *hyperplane*, for short, is a set associated with  $(s, r) \in \mathbb{R}^n \times \mathbb{R}$  ( $s \neq 0$ ) and defined by

$$H_{s,r} := \{x \in \mathbb{R}^n : \langle s, x \rangle = r\}.$$

An affine hyperplane is clearly a convex set. Fix  $s$  and let  $r$  describe  $\mathbb{R}$ ; then the affine hyperplanes  $H_{s,r}$  are translations of the same *linear*, or vector, hyperplane  $H_{s,0}$ . This  $H_{s,0}$  is the subspace of vectors that are orthogonal to  $s$  and can be denoted by  $H_{s,0} = \{s\}^\perp$ . Conversely, we say that  $s$  is the *normal* to  $H_{s,0}$  (up to a multiplicative constant). Affine hyperplanes play a fundamental role in convex analysis; the correspondence between  $0 \neq s \in \mathbb{R}^n$  and  $H_{s,1}$  is the basis for *duality* in a Euclidean space.

- (b) More generally, an *affine subspace*, or *affine manifold*, is a set  $V$  such that the (affine) line  $\{\alpha x + (1 - \alpha)x' : \alpha \in \mathbb{R}\}$  is entirely contained in  $V$  whenever  $x$  and  $x'$  are in  $V$  (note that a single point is an affine manifold). Again, an affine manifold is clearly convex.

Take  $v \in V$ ; it is easy – but instructive – to show that  $V - \{v\}$  is a subspace of  $\mathbb{R}^n$ , which is independent of the particular  $v$ ; denote it by  $V_0$ . Thus, an affine manifold  $V$  is nothing but the translation of some vector space  $V_0$ , sometimes called the *direction* (-subspace) of  $V$ . One can therefore speak of the *dimension* of an affin manifold  $V$ : it is just the dimension of  $V_0$ . We summarize in Table 1.1.1 the particular cases of affine manifolds.

**Table 1.1.1.** Various affine manifolds

Name	Possible definition	Direction	Dimension
point	$\{x\}$ ( $x \in \mathbb{R}^n$ )	$\{0\}$	0
affine line	$\{\alpha x_1 + (1 - \alpha)x_2 : \alpha \in \mathbb{R}\}$ $x_1 \neq x_2$ (both in $\mathbb{R}^n$ )	vector line $\mathbb{R}(x - x')$	1
affine hyperplane	$\{x \in \mathbb{R}^n : \langle s, x \rangle = r\}$ $(s \neq 0, r \in \mathbb{R})$	vector hyperpl. $\{s\}^\perp$	$n - 1$

- (c) The *half-spaces* of  $\mathbb{R}^n$  are those sets attached to  $(s, r) \in \mathbb{R}^n \times \mathbb{R}$ ,  $s \neq 0$ , and defined by

$$\begin{aligned} &\{x \in \mathbb{R}^n : \langle s, x \rangle \leq r\} \quad (\text{closed half-space}) \\ &\{x \in \mathbb{R}^n : \langle s, x \rangle < r\} \quad (\text{open half-space}); \end{aligned}$$

“affine half-space” would be a more accurate terminology. Naturally, an open [resp. closed] half-space is really an open [resp. closed] set; it is the interior [resp. closure] of the corresponding closed [resp. open] half-space; and the affine hyperplanes are the boundaries of the half-spaces; all this essentially comes from the continuity of the scalar product  $\langle s, \cdot \rangle$ .  $\square$

**Example 1.1.3 (Simplices)** Call  $\alpha = (\alpha_1, \dots, \alpha_k)$  the generic point of the space  $\mathbb{R}^k$ . The *unit simplex* in  $\mathbb{R}^k$  is

$$\Delta_k := \left\{ \alpha \in \mathbb{R}^k : \sum_{i=1}^k \alpha_i = 1, \alpha_i \geq 0 \text{ for } i = 1, \dots, k \right\}.$$

Equipping  $\mathbb{R}^k$  with the standard dot-product,  $\{e_1, \dots, e_k\}$  being the canonical basis and  $e := (1, \dots, 1)$  the vector whose coordinates are all 1, we can also write

$$\Delta_k := \left\{ \alpha \in \mathbb{R}^k : e^\top \alpha = 1, e_i^\top \alpha \geq 0 \text{ for } i = 1, \dots, k \right\}. \quad (1.1.1)$$

Observe the hyperplane and half-spaces appearing in this definition. Unit simplices are convex, compact, and have empty interior – being included in an affine hyperplane. We will often refer to a point  $\alpha \in \Delta_k$  as a set of ( $k$ ) *convex multipliers*.

It is sometimes useful to embed  $\Delta_k$  in  $\mathbb{R}^m$ ,  $m > k$ , by appending  $m - k$  zeros to the coordinates of  $\alpha \in \mathbb{R}^k$ , thus obtaining a vector of  $\Delta_m$ . We mention that a so-called *simplex* of  $\mathbb{R}^n$  is the figure formed by  $n + 1$  vectors in “nondegenerate positions”; in this sense, the unit simplex of  $\mathbb{R}^k$  is a simplex in the affine hyperplane of equation  $e^\top \alpha = 1$ ; see Fig. 1.1.1.

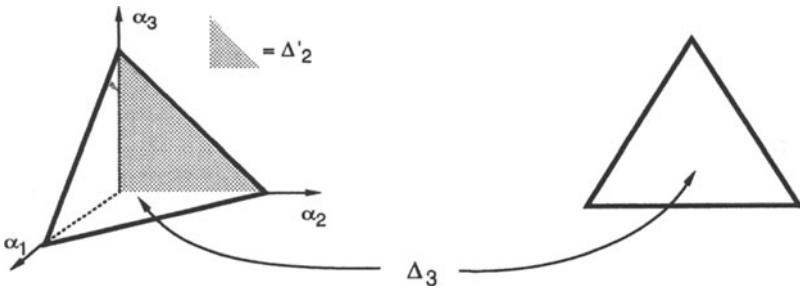


Fig. 1.1.1. Representing a simplex

If we replace  $e^\top \alpha = 1$  in (1.1.1) by  $e^\top \alpha \leq 1$ , we obtain another important set, convex, compact, with nonempty interior:

$$\Delta'_k := \left\{ \alpha \in \mathbb{R}^k : e^\top \alpha \leq 1, \alpha_i \geq 0 \text{ for } i = 1, \dots, k \right\}.$$

In fact, this set can also be described as follows:

$$\alpha \in \Delta'_k \iff \exists \alpha_{k+1} \geq 0 \text{ such that } (\alpha, \alpha_{k+1}) \in \Delta_{k+1}.$$

In this sense, the simplex  $\Delta'_k \subset \mathbb{R}^k$  can be identified with  $\Delta_{k+1}$  via a projection operator.

A (unit) simplex is traditionally visualized by a triangle, which can represent  $\Delta_3$  or  $\Delta'_2$ ; see Fig. 1.1.1 again.  $\square$

**Example 1.1.4 (Convex Cones)** A *cone*  $K$  is a set such that the “open” half-line  $\{\alpha x : \alpha > 0\}$  is entirely contained in  $K$  whenever  $x \in K$ . In the usual representation of geometrical objects, a cone has an apex; this apex is here at 0 (when it exists: a subspace is a cone but has no apex in this intuitive sense). Also,  $K$  is not supposed to contain 0 – this is mainly for notational reasons, to avoid writing  $0 \times (+\infty)$  in some situations. A *convex cone* is of course a cone which is convex; an example is the set defined in  $\mathbb{R}^n$  by

$$\langle s_j, x \rangle = 0 \text{ for } j = 1, \dots, m, \quad \langle s_{m+j}, x \rangle \leq 0 \text{ for } j = 1, \dots, p, \quad (1.1.2)$$

where the  $s_j$ 's are given in  $\mathbb{R}^n$  (once again, observe the hyperplanes and the half-spaces appearing in the above example, observe also that the defining relations must have zero right-hand sides).

Convexity of a given set is easier to check if this set is already known to be a cone: in view of Definition 1.1.1, a cone  $K$  is convex if and only if

$$x + y \in K \quad \text{whenever } x \text{ and } y \text{ are in } K,$$

i.e.  $K + K \subset K$ . Subspaces are particular convex cones. We leave it as an exercise to show that, to become a subspace, what is missing from a convex cone is just symmetry ( $K = -K$ ).

A very simple cone is the *nonnegative orthant* of  $\mathbb{R}^n$

$$\Omega_+ := \{x = (\xi^1, \dots, \xi^n) : \xi^i \geq 0 \text{ for } i = 1, \dots, n\}.$$

It can also be represented in terms of the canonical basis:

$$\Omega_+ = \left\{ \sum_{i=1}^n \alpha_i e_i : \alpha_i \geq 0 \text{ for } i = 1, \dots, n \right\}$$

or, in the spirit of (1.1.2):

$$\Omega_+ = \{x \in \mathbb{R}^n : \langle e_i, x \rangle \geq 0 \text{ for } i = 1, \dots, n\}.$$

Convex cones will be of fundamental use in the sequel, as they are among the simplest convex sets. Actually, they are important in convex analysis (the “unilateral” realm of inequalities), just as subspaces are important in linear analysis (the “bilateral” realm of equalities).  $\square$

## 1.2 Convexity-Preserving Operations on Sets

**Proposition 1.2.1** *Let  $\{C_j\}_{j \in J}$  be an arbitrary family of convex sets. Then*

$$C := \cap \{C_j : j \in J\}$$

*is convex.*

PROOF. Immediate from the very Definition 1.1.1.  $\square$

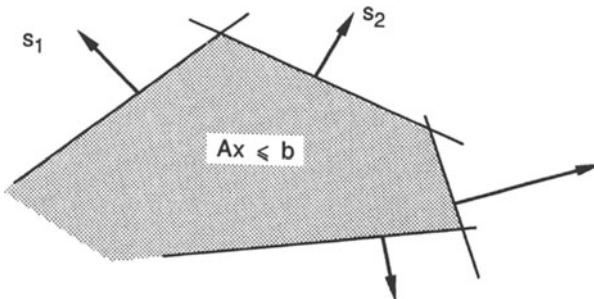
Intersecting convex sets is an operation of utmost importance; on the other hand, a union of convex sets is usually not convex.

**Example 1.2.2** Let  $(s_1, r_1), \dots, (s_m, r_m)$  be  $m$  given elements of  $\mathbb{R}^n \times \mathbb{R}$  and consider the set

$$\{x \in \mathbb{R}^n : \langle s_j, x \rangle \leq r_j \text{ for } j = 1, \dots, m\}. \quad (1.2.1)$$

It is clearly convex, which is confirmed if we view it as an intersection of  $m$  half-spaces; see Fig. 1.2.1.

We find it convenient to introduce two notations;  $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is the linear operator which, to  $x \in \mathbb{R}^n$ , associates the vector with coordinates  $\langle s_j, x \rangle$ ; and in  $\mathbb{R}^m$ , the notation  $a \leq b$  means that each coordinate of  $a$  is lower than or equal to the corresponding coordinate of  $b$ . Then, the set (1.2.1) can be characterized by  $Ax \leq b$ , where  $b \in \mathbb{R}^m$  is the vector with coordinates  $r_1, \dots, r_m$ .  $\square$



**Fig. 1.2.1.** An intersection of half-spaces

It is interesting to observe that the above construction applies to the examples of §1.1:

- an affine hyperplane is the intersection of two (closed) half-spaces;
- an affine manifold is an intersection of finitely many affine hyperplanes;
- a unit simplex is the intersection of an affine hyperplane with a closed convex cone ( $\Omega_+$ );
- a convex cone such as in (1.1.2) is an intersection of a subspace with (homogeneous) half-spaces.

Piecing together these instances of convex sets, we see that they can all be considered as intersections of sufficiently many closed half-spaces. Another observation is that, up to a translation, a hyperplane is the simplest instance of a convex cone – apart from (linear) subspaces. Conclusion: *translations* (the key operations in the affine world), *intersections* and closed *half-spaces* are basic objects in convex analysis.

Convexity is stable under Cartesian product, just as it is under intersection.

**Proposition 1.2.3** *For  $i = 1, \dots, k$ , let  $C_i \subset \mathbb{R}^{n_i}$  be convex sets. Then  $C_1 \times \dots \times C_k$  is a convex set of  $\mathbb{R}^{n_1} \times \dots \times \mathbb{R}^{n_k}$ .*

PROOF. Straightforward. □

The converse is also true;  $C_1 \times \dots \times C_k$  is convex if and only if each  $C_i$  is convex, and this results from the next property: stability under affine mappings. We recall that  $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is said *affine* when

$$A(\alpha x + (1 - \alpha)x') = \alpha A(x) + (1 - \alpha)A(x')$$

for all  $x$  and  $x'$  in  $\mathbb{R}^n$  and all  $\alpha \in \mathbb{R}$ . This means that  $x \mapsto A(x) - A(0)$  is linear, so an affine mapping can be characterized by a linear mapping  $A_0$  and a point  $y_0 := A(0) \in \mathbb{R}^m$ :

$$A(x) = A_0x + y_0 \quad \text{for all } x \in \mathbb{R}^n.$$

It goes without saying that images of affine manifolds under affine mappings are affine manifolds (hence the name!) So is the case as well for convex sets:

**Proposition 1.2.4** Let  $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$  be an affine mapping and  $C$  a convex set of  $\mathbb{R}^n$ . The image  $A(C)$  of  $C$  under  $A$  is convex in  $\mathbb{R}^m$ .

If  $D$  is a convex set of  $\mathbb{R}^m$ , the inverse image

$$A^{-1}(D) := \{x \in \mathbb{R}^n : A(x) \in D\}$$

is convex in  $\mathbb{R}^n$ .

PROOF. For  $x$  and  $x'$  in  $\mathbb{R}^n$ , the image under  $A$  of the segment  $[x, x']$  is clearly the segment  $[A(x), A(x')] \subset \mathbb{R}^m$ . This proves the first claim, but also the second: indeed, if  $x$  and  $x'$  are such that  $A(x)$  and  $A(x')$  are both in the convex set  $D$ , then every point of the segment  $[x, x']$  has its image in  $[A(x), A(x')] \subset D$ .  $\square$

Immediate consequences of this last result are:

- the opposite  $-C$  of a convex set is convex;
- the sum (called direct sum, or *Minkowski* sum, denoted with the symbol  $\oplus$  by some authors)

$$C_1 + C_2 := \{x = x_1 + x_2 : x_1 \in C_1, x_2 \in C_2\}$$

of two convex sets  $C_1$  and  $C_2$  is convex; when  $C_2 = \{c_2\}$  is a singleton, we will sometimes use the lighter notation  $C_1 + c_2$  for  $C_1 + \{c_2\}$ ;

- more generally, if  $\alpha_1$  and  $\alpha_2$  are two real numbers, the set

$$\alpha_1 C_1 + \alpha_2 C_2 := \{\alpha_1 x_1 + \alpha_2 x_2 : x_1 \in C_1, x_2 \in C_2\} \quad (1.2.2)$$

is convex: it is the image of the convex set  $C_1 \times C_2$  (Proposition 1.2.3) under the linear mapping sending  $(x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n$  to  $\alpha_1 x_1 + \alpha_2 x_2 \in \mathbb{R}^n$ .

We recall here that the sum of two closed sets need not be closed, unless one of the sets is compact. This property is not changed when convexity is present: with  $n = 2$ , take for example

$$C_1 := \{(\xi, \eta) : \xi \geq 0, \eta \geq 0, \xi\eta \geq 1\} \quad \text{and} \quad C_2 := \mathbb{R} \times \{0\}.$$

**Example 1.2.5** Let  $C$  be convex in  $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ ; see Fig. 1.2.2.

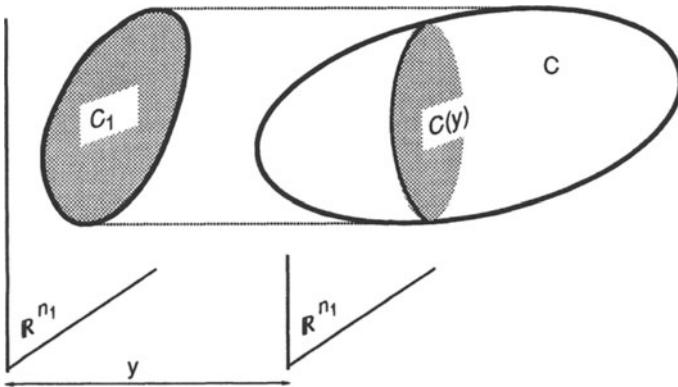
Use for  $A$  the projection from  $\mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$  onto  $\mathbb{R}^{n_1}$  to see that the “slice” of  $C$  along  $y$

$$C(y) := \{x \in \mathbb{R}^{n_1} : (x, y) \in C\}$$

and the “shadow” of  $C$  over  $\mathbb{R}^{n_1}$

$$C_1 := \{x \in \mathbb{R}^{n_1} : (x, y) \in C \text{ for some } y \in C\}$$

are convex. If, in particular,  $C = C_1 \times C_2$  is a product-set, we obtain the converse to Proposition 1.2.3.  $\square$



**Fig. 1.2.2.** Shadow and slice of a convex set

**Example 1.2.6** When setting  $\alpha_1 = -\alpha_2 = 1$  in (1.2.2), we obtain a “difference”  $C_1 - C_2$ , which is actually a sum:  $C_1 + (-C_2)$ ; the result is always a rather large set, as it contains “as many elements” as  $C_1 \times C_2$ , so to speak.

Another “difference” between sets is the following *star-difference*

$$C_1 \pm C_2 := \cap \{C_1 - c : c \in C_2\} = \{x \in \mathbb{R}^n : x + C_2 \subset C_1\}.$$

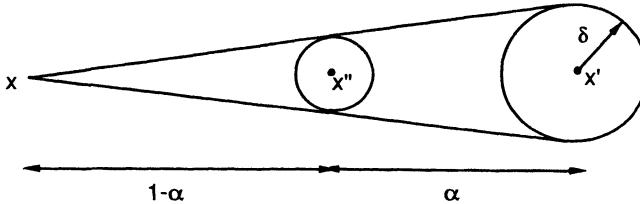
It is a convex set, even if  $C_2$  is not convex – use in particular Proposition 1.2.1. Observe the contrast:  $C_1 - C_2$  is a “large” set but  $C_1 \pm C_2$  is “small”, and even very often empty (try examples with a ball for  $C_2$ ).  $\square$

We finish with a topological operation.

**Proposition 1.2.7** *If  $C$  is convex, so are its interior  $\text{int } C$  and its closure  $\text{cl } C$ .*

PROOF. For given different  $x$  and  $x'$ , and  $\alpha \in ]0, 1[$ , we set  $x'' = \alpha x + (1 - \alpha)x' \in ]x, x'[$ .

Take first  $x$  and  $x'$  in  $\text{int } C$ . Choosing  $\delta > 0$  such that  $B(x', \delta) \subset C$ , we show that  $B(x'', (1 - \alpha)\delta) \subset C$ . As often in convex analysis, it is probably best to draw a picture. The ratio  $\|x'' - x\|/\|x' - x\|$  being precisely  $1 - \alpha$ , Fig. 1.2.3 clearly shows that  $B(x'', (1 - \alpha)\delta)$  is just the set  $\alpha x + (1 - \alpha)B(x', \delta)$ , obtained from segments with endpoints in  $\text{int } C$ :  $x'' \in \text{int } C$ .



**Fig. 1.2.3.** Convex sets have convex interiors

Now, take  $x$  and  $x'$  in  $\text{cl } C$ : we select in  $C$  two sequences  $\{x_k\}$  and  $\{x'_k\}$  converging to  $x$  and  $x'$  respectively. Then,  $\alpha x_k + (1 - \alpha)x'_k$  is in  $C$  and converges to  $x''$ , which is therefore in  $\text{cl } C$ .  $\square$

The interior of a set is (too) often empty; convexity allows the similar but much more convenient concept of *relative interior*, to be seen below in §2.1. Observe the nonsymmetric character of  $x$  and  $x'$  in Fig. 1.2.3. It can be exploited to show that the intermediate result  $[x, x'[\subset \text{int } C$  remains true even if  $x \in \text{cl } C$ ; a property which will be seen in more detail in §2.1.

### 1.3 Convex Combinations and Convex Hulls

The operations described in §1.2 took convex sets and made new convex sets with them. The present section is devoted to another operation, which takes a nonconvex set and makes a convex set with it. First, let us recall the following basic facts from linear algebra.

- (i) A *linear combination* of elements  $x_1, \dots, x_k$  of  $\mathbb{R}^n$  is an element  $\sum_{i=1}^k \alpha_i x_i$ , where the coefficients  $\alpha_i$  are arbitrary real numbers.
- (ii) A (linear) *subspace* of  $\mathbb{R}^n$  is a set containing all its linear combinations; an intersection of subspaces is still a subspace.
- (iii) To any nonempty set  $S \subset \mathbb{R}^n$ , we can therefore associate the intersection of all subspaces containing  $S$ . This gives a subspace: the subspace generated by  $S$  (or *linear hull* of  $S$ ), denoted  $\text{lin } S$  – other notations are  $\text{vect } S$  or  $\text{span } S$ .
- (iv) For the  $\subset$ -relation,  $\text{lin } S$  is the smallest subspace containing  $S$ ; it can be constructed directly from  $S$ , by collecting all the linear combinations of elements of  $S$ .
- (v) Finally,  $x_1, \dots, x_k$  are said linearly independent if  $\sum_{i=1}^k \alpha_i x_i = 0$  implies that  $\alpha_1 = \dots = \alpha_k = 0$ . In  $\mathbb{R}^n$ , this implies  $k \leq n$ .

Now, let us be slightly more demanding for the coefficients  $\alpha_i$ , as follows:

- (i') An *affine combination* of elements  $x_1, \dots, x_k$  of  $\mathbb{R}^n$  is an element  $\sum_{i=1}^k \alpha_i x_i$ , where the coefficients  $\alpha_i$  satisfy  $\sum_{i=1}^k \alpha_i = 1$ .

As explained after Example 1.2.2, “affinity = linearity + translation”, it is therefore not surprising to realize that the development (i) – (v) can be reproduced starting from (i'):

- (ii') An affine manifold in  $\mathbb{R}^n$  is a set containing all its affine combinations (the equivalence with Example 1.1.2(b) will appear more clearly below in Proposition 1.3.3); it is easy to see that an intersection of affine manifolds is still an affine manifold.
- (iii') To any nonempty set  $S \subset \mathbb{R}^n$ , we can therefore associate the intersection of all affine manifolds containing  $S$ . This gives the affine manifold generated by  $S$ , denoted  $\text{aff } S$ : the *affine hull* of  $S$ .

(iv') For the  $\subset$ -relation,  $\text{aff } S$  is the smallest affine manifold containing  $S$ ; it can be constructed directly from  $S$ , by collecting all the affine combinations of elements of  $S$ . To see it, start from  $x_0 \in S$ , take  $\text{lin}(S - x_0)$  and come back by adding  $x_0$ : the result  $x_0 + \text{lin}(S - x_0)$  is just  $\text{aff } S$ .

(v') Finally, the  $k + 1$  points  $x_0, x_1, \dots, x_k$  are said affinely independent if the set

$$x_0 + \text{lin}\{x_0 - x_0, x_1 - x_0, \dots, x_k - x_0\} = x_0 + \text{lin}\{x_1 - x_0, \dots, x_k - x_0\}$$

has full dimension, namely  $k$ . The above set is exactly  $\text{aff}\{x_0, x_1, \dots, x_k\}$ ; hence, it does not depend on the index chosen for the translation (here 0). In linear language, the required property is that the  $k$  vectors  $x_i - x_0, i \neq 0$  be linearly independent. Getting rid of the arbitrary index 0, this means that the system of equations

$$\sum_{i=0}^k \alpha_i x_i = 0, \quad \sum_{i=0}^k \alpha_i = 0 \quad (1.3.1)$$

has the unique solution  $\alpha_0 = \alpha_1 = \dots = \alpha_k = 0$ . Considered as elements of  $\mathbb{R}^{n+1} = \mathbb{R}^n \times \mathbb{R}$ , the vectors  $(x_0, 1), (x_1, 1), \dots, (x_k, 1)$ , are linearly independent. In  $\mathbb{R}^n$ , at most  $n + 1$  elements can thus be affinely independent.

If  $x_0, x_1, \dots, x_k$ , are affinely independent,  $x \in \text{aff}\{x_0, x_1, \dots, x_k\}$  can be written in a unique way as

$$x = \sum_{i=0}^k \alpha_i x_i \quad \text{with} \quad \sum_{i=0}^k \alpha_i = 1.$$

The corresponding coefficients  $\alpha_i$  are sometimes called the *barycentric coordinates* of  $x$  – even though such a terminology should be reserved to nonnegative  $\alpha_i$ 's. To say that a set of vectors are affinely dependent is to say that one of them (any one) is an affine combination of the others.

**Example 1.3.1** Consider the unit simplex  $\Delta_3$  on the left part of Fig. 1.1.1; call  $e_1 = (1, 0, 0)$ ,  $e_2 = (0, 1, 0)$ ,  $e_3 = (0, 0, 1)$  the three basis-vectors forming its vertices. The affine hull of  $S = \{e_1, e_2\}$  is the affine line passing through  $e_1$  and  $e_2$ . For  $S = \{e_1, e_2, e_3\}$ , it is the affine plane of equation  $\alpha_1 + \alpha_2 + \alpha_3 = 1$ . The four elements  $0, e_1, e_2, e_3$  are affinely independent but the four elements  $(1/3, 1/3, 1/3), e_1, e_2, e_3$  are not.  $\square$

Passing from (i) to (i') gives a set  $\text{aff } S$  which is closer to  $S$  than  $\text{lin } S$ , thanks to the extra requirement in (i'). We apply once more the same idea and we pass from affinity to convexity by requiring some more of the  $\alpha_i$ 's. This gives a new definition, playing the role of (i) and (i'):

**Definition 1.3.2** A *convex combination* of elements  $x_1, \dots, x_k$  in  $\mathbb{R}^n$  is an element of the form

$$\sum_{i=1}^k \alpha_i x_i \quad \text{where} \quad \sum_{i=1}^k \alpha_i = 1 \quad \text{and} \quad \alpha_i \geq 0 \text{ for } i = 1, \dots, k. \quad \square$$

A convex combination is therefore a particular affine combination, which in turn is a particular linear combination. Note in passing that all convex combinations of given  $x_1, \dots, x_k$  form a convex set: it is the image of  $\Delta_k$  under the linear mapping

$$\mathbb{R}^k \ni (\alpha_1, \dots, \alpha_k) \mapsto \alpha_1 x_1 + \dots + \alpha_k x_k \in \mathbb{R}^n.$$

The sets playing the role of linear or affine subspaces of (ii) and (ii') will now be logically called convex, but we have to make sure that this new definition is consistent with Definition 1.1.1.

**Proposition 1.3.3** *A set  $C \subset \mathbb{R}^n$  is convex if and only if it contains every convex combination of its elements.*

PROOF. The condition is sufficient: convex combinations of two elements just make up the segment joining them. To prove necessity, take  $x_1, \dots, x_k$  in  $C$  and  $\alpha = (\alpha_1, \dots, \alpha_k) \in \Delta_k$ . One at least of the  $\alpha_i$ 's is positive, say  $\alpha_1 > 0$ . Then form

$$y_2 := \frac{\alpha_1}{\alpha_1 + \alpha_2} x_1 + \frac{\alpha_2}{\alpha_1 + \alpha_2} x_2 \quad \left[ = \frac{1}{\alpha_1 + \alpha_2} (\alpha_1 x_1 + \alpha_2 x_2) \right]$$

which is in  $C$  by Definition 1.1.1 itself. Therefore,

$$y_3 := \frac{\alpha_1 + \alpha_2}{\alpha_1 + \alpha_2 + \alpha_3} y_2 + \frac{\alpha_3}{\alpha_1 + \alpha_2 + \alpha_3} x_3 \quad \left[ = \frac{1}{\sum_{i=1}^3 \alpha_i} \sum_{i=1}^3 \alpha_i x_i \right]$$

is in  $C$  for the same reason; and so on until

$$y_k := \frac{\alpha_1 + \dots + \alpha_{k-1}}{1} y_{k-1} + \frac{\alpha_k}{1} x_k \quad \left[ = \frac{1}{1} \sum_{i=1}^k \alpha_i x_i \right]. \quad \square$$

The working argument of the above proof is longer to write than to understand. Its basic idea is just *associativity*: a convex combination  $x = \sum \alpha_i x_i$  of convex combinations  $x_i = \sum \beta_{ij} y_{ij}$  is still a convex combination  $x = \sum \sum (\alpha_i \beta_{ij}) y_{ij}$ . The same associativity property will be used in the next result.

Because an intersection of convex sets is convex, we can logically define as in (iii), (iii') the *convex hull*  $\text{co } S$  of a nonempty set  $S$ : this is the intersection of all the convex sets containing  $S$ .

**Proposition 1.3.4** *The convex hull can also be described as the set of all convex combinations:*

$$\begin{aligned} \text{co } S &:= \cap \{C : C \text{ is convex and contains } S\} \\ &= \left\{ x \in \mathbb{R}^n : \text{for some } k \in \mathbb{N}_*, \text{ there exist } x_1, \dots, x_k \in S \text{ and } \alpha = (\alpha_1, \dots, \alpha_k) \in \Delta_k \text{ such that } \sum_{i=1}^k \alpha_i x_i = x \right\}. \end{aligned} \quad (1.3.2)$$

PROOF. Call  $T$  the set described in the rightmost side of (1.3.2). Clearly,  $T \supset S$ . Also, if  $C$  is convex and contains  $S$ , then it contains all convex combinations of elements

in  $S$  (Proposition 1.3.3), i.e.  $C \supset T$ . The proof will therefore be finished if we show that  $T$  is convex.

For this, take two points  $x$  and  $y$  in  $T$ , characterized respectively by  $(x_1, \alpha_1), \dots, (x_k, \alpha_k)$  and by  $(y_1, \beta_1), \dots, (y_\ell, \beta_\ell)$ ; take also  $\lambda \in ]0, 1[$ . Then  $\lambda x + (1 - \lambda)y$  is a certain combination of  $k + \ell$  elements of  $S$ ; this combination is convex because its coefficients  $\lambda\alpha_i$  and  $(1 - \lambda)\beta_j$  are nonnegative, and their sum is

$$\lambda \sum_{i=1}^k \alpha_i + (1 - \lambda) \sum_{j=1}^\ell \beta_j = \lambda + 1 - \lambda = 1. \quad \square$$

**Example 1.3.5** Take a finite set  $\{x_1, \dots, x_m\}$ . To obtain its convex hull, it is not necessary to list all the convex combinations obtained via  $\alpha \in \Delta_k$  for all  $k = 1, \dots, m$ . In fact, as already seen in Example 1.1.3,  $\Delta_k \subset \Delta_m$  if  $k \leq m$ , so we can restrict ourselves to  $k = m$ . Thus, we see that

$$\text{co}\{x_1, \dots, x_m\} = \left\{ \sum_{j=1}^m \alpha_j x_j : \alpha = (\alpha_1, \dots, \alpha_m) \in \Delta_m \right\}.$$

Make this example a little more complicated, replacing the collection of points by a collection of convex sets:

$$S = C_1 \cup \dots \cup C_m \quad \text{where each } C_i \text{ is convex.}$$

A simplification of (1.3.2) can again be exploited here. Indeed, consider a convex combination  $\sum_{i=1}^k \alpha_i x_i$ . It may happen that several of the  $x_i$ 's belong to the same  $C_j$ . To simplify notation, suppose that  $x_{k-1}$  and  $x_k$  are in  $C_1$ ; assume also  $\alpha_k > 0$ . Then set  $(\beta_i, y_i) := (\alpha_i, x_i)$ ,  $i = 1, \dots, k-2$  and

$$\beta_{k-1} := \alpha_{k-1} + \alpha_k, \quad y_{k-1} := \frac{1}{\beta_{k-1}}(\alpha_{k-1}x_{k-1} + \alpha_k x_k) \in C_1,$$

so that  $\sum_{i=1}^k \alpha_i x_i = \sum_{i=1}^{k-1} \beta_i y_i$ . Our convex combination  $(\alpha, x)$  is useless, in the sense that it can also be found among those with  $k-1$  elements. To cut a long story short, associativity of convex combinations yields

$$\text{co } S = \left\{ \sum_{i=1}^m \alpha_i x_i : \alpha \in \Delta_m, \quad x_i \in C_i \text{ for } i = 1, \dots, m \right\}.$$

From a geometrical point of view, the convex hull of  $C_1 \cup C_2$  ( $m = 2$ ) is simply constructed by drawing segments, with endpoints in  $C_1$  and  $C_2$ ; for  $C_1 \cup C_2 \cup C_3$ , we paste triangles, etc.  $\square$

When  $S$  is infinite, or has infinitely many convex components,  $k$  is a priori unbounded in (1.3.2) and cannot be readily restricted as in the examples above. Yet, a bound on  $k$  exists for all  $S$  when we consider linear combinations and linear hulls – and consequently in the affine case as well; this is the whole business of *dimension*. In the present case of convex combinations, the same phenomenon is conserved to some extent. For each positive integer  $k$ , call  $S_k$  the set of all convex combinations of  $k$  elements in  $S$ : we have

$$S = S_1 \subset S_2 \subset \cdots \subset S_k \subset \cdots$$

The  $S_k$ 's are not convex but, “at the limit”, their union is convex and coincides with  $\text{co } S$  (Proposition 1.3.4). The theorem below tells us that  $k$  does not have to go to  $+\infty$ : the above sequence actually stops at  $S_{n+1} = \text{co } S$ .

**Theorem 1.3.6 (C. Carathéodory)** *Any  $x \in \text{co } S \subset \mathbb{R}^n$  can be represented as a convex combination of  $n + 1$  elements of  $S$ .*

PROOF. Take an arbitrary convex combination  $x = \sum_{i=1}^k \alpha_i x_i$ , with  $k > n + 1$ . We claim that one of the  $x_i$ 's can be assigned a 0-coefficient without changing  $x$ . For this, assume that all coefficients  $\alpha_i$  are positive (otherwise we are done).

The  $k > n + 1$  elements  $x_i$  are certainly affinely dependent: (1.3.1) tells us that we can find  $\delta_1, \dots, \delta_k$ , not all zero, such that

$$\sum_{i=1}^k \delta_i x_i = 0 \quad \text{and} \quad \sum_{i=1}^k \delta_i = 0.$$

There is at least one positive  $\delta_i$  and we can set  $\alpha'_i := \alpha_i - t^* \delta_i$  for  $i = 1, \dots, k$ , where

$$t^* := \max \{t \geq 0 : \alpha_i - t\delta_i \geq 0 \text{ for } i = 1, \dots, k\} = \min_{\delta_j > 0} \frac{\alpha_j}{\delta_j}.$$

Clearly enough,

$$\begin{aligned} \alpha'_i &\geq 0 \quad \text{for } i = 1, \dots, k && \begin{aligned} &\text{[automatic if } \delta_i \leq 0, \\ &\text{by construction of } t^* \text{ if } \delta_i > 0] \end{aligned} \\ \sum_{i=1}^k \alpha'_i &= \sum_{i=1}^k \alpha_i - t^* \sum_{i=1}^k \delta_i = 1; \\ \sum_{i=1}^k \alpha'_i x_i &= x - t^* \sum_{i=1}^k \delta_i x_i = x; \\ \exists i_0 &\quad \text{such that } \alpha'_{i_0} = 0. && \text{[by construction of } t^*] \end{aligned}$$

In other words, we have expressed  $x$  as a convex combination of  $k - 1$  among the  $x_i$ 's; our claim is proved.

Now, if  $k - 1 = n + 1$ , the proof is finished. If not, we can apply the above construction to the convex combination  $x = \sum_{i=1}^{k-1} \alpha'_i x_i$  and so on. The process can be continued until there remain only  $n + 1$  elements (which may be affinely independent).  $\square$

The same proof technique is commonly used in actual computations dealing with linearly constrained optimization. Geometrically, we start from  $\alpha = (\alpha_1, \dots, \alpha_k) \in \Delta_k$ . We compute a direction  $-d = (\delta_1, \dots, \delta_k)$ , which is in the subspace parallel to  $\text{aff } \Delta_k$ , so that for any stepsize  $t$ ,  $\alpha - td \in \text{aff } \Delta_k$ ; and also,  $x$  is kept invariant. The particular  $t^*$  is the maximal stepsize such that  $\alpha - td \in \Delta_k$ ; as a result,  $\alpha - t^*d$  is on the boundary of  $\Delta_k$ , i.e. in  $\Delta_{k-1}$ ; see Fig. 1.3.1.

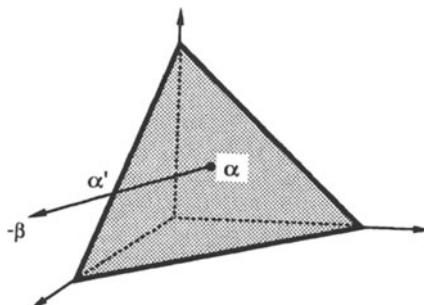


Fig. 1.3.1. Carathéodory's theorem

The theorem of Carathéodory does not establish the existence of a “basis” with  $n + 1$  elements, as is the case for linear combinations. Here, the generators  $x_i$  may depend on the particular  $x$  to be computed. In  $\mathbb{R}^2$ , think of the corners of a square: any one of these  $4 > 2 + 1$  points may be necessary to generate a point in the square; also, the unit disk cannot be generated by finitely many points on the unit circle. By contrast, a subspace of dimension  $m$  can be generated by  $m$  (carefully selected but) *fixed* generators.

It is not the particular value  $n + 1$  which is interesting in the above theorem, but rather the fact that the cardinality of relevant convex combinations is bounded: this is particularly useful when passing to the limit in a sequence of convex combinations. This value  $n + 1$  is not of fundamental importance, anyway, and can often be reduced – as in Example 1.3.5: the convex hull of two convex sets in  $\mathbb{R}^{100}$  can be generated by 2-combinations; also, the technique of proof shows that it is the dimension of  $\text{aff } S$  that counts, not  $n$ . Along these lines, we mention without proof a result geometrically very suggestive:

**Theorem 1.3.7 (W. Fenchel and L. Bunt)** *If  $S \subset \mathbb{R}^n$  has no more than  $n$  connected components (in particular, if  $S$  is connected), then any  $x \in \text{co } S$  can be expressed as a convex combination of  $n$  elements of  $S$ .*  $\square$

This result says in particular that convex and connected one-dimensional sets are the same, namely the intervals. In  $\mathbb{R}^2$ , the convex hull of a continuous curve can be obtained by joining all pairs of points in it. In  $\mathbb{R}^3$ , the convex hull of three potatoes is obtained by pasting triangles, etc.

## 1.4 Closed Convex Sets and Hulls

Closedness is a very important property in convex analysis and optimization. Most of the convex sets of interest to us in the subsequent chapters will be closed. It is therefore relevant to reproduce the previous section, with the word “closed” added. As far as linearity and affinity are concerned, there is no difference; in words, equalities are not affected when limits are involved. But convexity is another story: when passing from (i), (i') to Definition 1.3.2, inequalities are introduced, together with their accompanying difficulty “ $<$  vs.  $\leqslant$ ”.

To construct a convex hull  $\text{co } S$ , we followed in §1.3 the path (iii), (iii'): we took the intersection of all convex sets containing  $S$ . An intersection of closed sets is still closed, so the following definition is also natural:

**Definition 1.4.1** The *closed convex hull* of a nonempty set  $S \subset \mathbb{R}^n$  is the intersection of all closed convex sets containing  $S$ . It will be denoted by  $\overline{\text{co}} S$ .  $\square$

Another path was also possible to construct  $\text{co } S$ , namely to take all possible convex combinations: then, we obtained  $\text{co } S$  again (Proposition 1.3.4); what about closing it? It turns out we can do that as well:

**Proposition 1.4.2** *The closed convex hull  $\overline{\text{co}} S$  of Definition 1.4.1 is the closure  $\text{cl}(\text{co } S)$  of the convex hull of  $S$ .*

PROOF. Because  $\text{cl}(\text{co } S)$  is a closed convex set containing  $S$ , it contains  $\overline{\text{co}} S$  as well. On the other hand, take a closed convex set  $C$  containing  $S$ ; being convex,  $C$  contains  $\text{co } S$ ; being closed, it contains also the closure of  $\text{co } S$ . Since  $C$  was arbitrary, we conclude  $\cap C \supset \text{cl co } S$ .  $\square$

From the very definitions, the operation “taking a hull” is monotone: if  $S_1 \subset S_2$ , then  $\text{aff } S_1 \subset \text{aff } S_2$ ,  $\text{cl } S_1 \subset \text{cl } S_2$ ,  $\text{co } S_1 \subset \text{co } S_2$ , and of course  $\overline{\text{co}} S_1 \subset \overline{\text{co}} S_2$ . A closed convex hull does not distinguish a set from its closure, just as it does not distinguish it from its convex hull:  $\overline{\text{co}} S = \overline{\text{co}}(\text{cl } S) = \overline{\text{co}}(\text{co } S)$ .

When computing  $\overline{\text{co}}$  via Proposition 1.4.2, the closure operation is necessary ( $\text{co } S$  need not be closed) and must be performed *after* taking the convex hull: the operations do not commute. Consider the example of Fig. 1.4.1:

$$S = \{(0, 0)\} \cup \{(\xi, 1) : \xi \geq 0\}.$$

It is a closed set but  $\text{co } S$  fails to be closed: it misses the half-line  $(\mathbb{R}^+, 0)$ . Nevertheless, this phenomenon can occur only when  $S$  is unbounded, a result which comes directly from Carathéodory’s theorem:

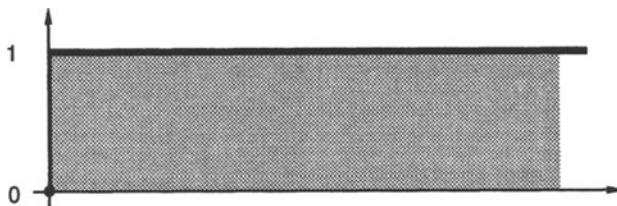


Fig. 1.4.1. A convex hull need not be closed

**Theorem 1.4.3** *If  $S$  is bounded [resp. compact], then  $\text{co } S$  is bounded [resp. compact].*

PROOF. Let  $x = \sum_{i=1}^{n+1} \alpha_i x_i \in \text{co } S$ . If  $S$  is bounded, say by  $M$ , we can write

$$\|x\| \leq \sum_{i=1}^{n+1} \alpha_i \|x_i\| \leq M \sum_{i=1}^{n+1} \alpha_i = M.$$

Now take a sequence  $\{x^k\} \subset \text{co } S$ . For each  $k$  we can choose

$$x_1^k, \dots, x_{n+1}^k \text{ in } S \quad \text{and} \quad \alpha^k = (\alpha_1^k, \dots, \alpha_{n+1}^k) \in \Delta_{n+1}$$

such that  $x^k = \sum_{i=1}^{n+1} \alpha_i^k x_i^k$ . Note that  $\Delta_{n+1}$  is compact. If  $S$  is compact, we can extract a subsequence as many times as necessary (not more than  $n+2$  times) so that  $\{\alpha^k\}$  and each  $\{x_i^k\}$  converge: we end up with an index set  $K \subset \mathbb{N}$  such that, when  $k \rightarrow +\infty$ ,

$$\{x_i^k\}_{k \in K} \rightarrow x_i \in S \quad \text{and} \quad \{\alpha^k\}_{k \in K} \rightarrow \alpha \in \Delta_{n+1}.$$

Passing to the limit for  $k \in K$ , we see that  $\{x^k\}_{k \in K}$  converges to a point  $x$ , which can be expressed as a convex combination of points of  $S$ :  $x \in \text{co } S$ , whose compactness is thus established.  $\square$

Thus, this theorem does allow us to write:

$$S \text{ bounded in } \mathbb{R}^n \implies \overline{\text{co}} S = \text{cl co } S = \text{co cl } S.$$

**Remark 1.4.4** Let us emphasize one point made clear by this and the previous sections: a hull (linear, affine, convex or closed) can be constructed in two ways. In the *inner* way, combinations (linear, affine, convex, or limits) are made with points taken from inside the starting set  $S$ . The *outer* way takes sets (linear, affine, convex, or closed) containing  $S$  and intersects them.

Even though the first way may seem more direct and natural, it is the second which must often be preferred, at least when convexity is involved. This is especially true when taking the closed convex hull: forming all convex combinations is already a nasty task, which is not even sufficient, as one must close the result afterwards. On the other hand, the external construction of  $\overline{\text{co}} S$  is more handy in a set-theoretic framework. We will even see in §4.2(b) that it is not necessary to take in Definition 1.4.1 all closed convex sets containing  $S$ : only rather special such sets have to be intersected, namely the closed half-spaces of Example 1.1.2(c).  $\square$

To finish this section, we mention one more hull, often useful. When starting from linear combinations to obtain convex combinations in Definition 1.3.2, we introduced two kinds of constraints on the coefficients:  $e^\top \alpha = 1$  and  $\alpha_i \geq 0$ . The first constraint alone yielded affinity; we can take the second alone:

**Definition 1.4.5** A *conical combination* of elements  $x_1, \dots, x_k$  is an element of the form  $\sum_{i=1}^k \alpha_i x_i$ , where the coefficients  $\alpha_i$  are nonnegative.

The set of all conical combinations from a given nonempty  $S \subset \mathbb{R}^n$  is the *conical hull* of  $S$ . It is denoted by  $\text{cone } S$ .  $\square$

Note that it would be more accurate to speak of convex conical combinations and convex conical hulls. If  $\bar{\alpha} := \sum_{i=1}^k \alpha_i$  is positive, we can set  $\beta_i := \alpha_i / \bar{\alpha}$  to realize that a conical combination of the type

$$\sum_{i=1}^k \alpha_i x_i = \bar{\alpha} \sum_{i=1}^k \beta_i x_i \quad \text{with} \quad \bar{\alpha} > 0, \quad \beta \in \Delta_k$$

is then nothing but a convex combination, multiplied by an arbitrary positive coefficient. We leave it to the reader to realize that

$$\text{cone } S = \mathbb{R}^+(\text{co } S) = \text{co}(\mathbb{R}^+ S).$$

Thus,  $0 \in \text{cone } S$ ; actually, to form  $\text{cone } S$ , we intersect all convex cones containing  $S$ , and we append  $0$  to the result. If we close it, we obtain the following definition:

**Definition 1.4.6** The *closed conical hull* (or rather closed convex conical hull) of a nonempty set  $S \subset \mathbb{R}^n$  is

$$\overline{\text{cone } S} := \text{cl cone } S = \text{cl} \left\{ \sum_{i=1}^k \alpha_i x_i : \alpha_i \geq 0, x_i \in S \text{ for } i = 1, \dots, k \right\}. \quad \square$$

Theorem 1.4.3 states that the convex hull and closed convex hull of a compact set coincide, but the property is no longer true for conical hulls: for a counter-example, take the set  $\{(\xi, \eta) \in \mathbb{R}^2 : (\xi - 1)^2 + \eta^2 \leq 1\}$ . Nevertheless, the result can be recovered with an additional assumption:

**Proposition 1.4.7** Let  $S$  be a nonempty compact set such that  $0 \notin \text{co } S$ . Then

$$\overline{\text{cone } S} = \mathbb{R}^+(\text{co } S) [= \text{cone } S].$$

PROOF. The set  $C := \text{co } S$  is compact and does not contain the origin; we prove that  $\mathbb{R}^+ C$  is closed. Let  $\{t_k x_k\} \subset \mathbb{R}^+ C$  converge to  $y$ ; extracting a subsequence if necessary, we may suppose  $x_k \rightarrow x \in C$ ; note:  $x \neq 0$ . We write

$$t_k \frac{x_k}{\|x_k\|} \rightarrow \frac{y}{\|x\|},$$

which implies  $t_k \rightarrow \|y\|/\|x\| =: t \geq 0$ . Then,  $t_k x_k \rightarrow t x = y$ , which is thus in  $\mathbb{R}^+ C$ .

$\square$

## 2 Convex Sets Attached to a Convex Set

### 2.1 The Relative Interior

Let  $C$  be a nonempty convex set in  $\mathbb{R}^n$ . If  $\text{int } C \neq \emptyset$ , one easily checks that the affine hull  $\text{aff } C$  is the whole of  $\mathbb{R}^n$  (because so is the affine hull of a ball contained in  $C$ ): we are dealing with a “full dimensional” set. On the other hand, let  $C$  be the sheet of paper on which this text is written. Its interior is empty in the surrounding space  $\mathbb{R}^3$ , but not in the space  $\mathbb{R}^2$  of the table on which it is lying; by contrast, note that  $\text{cl } C$  is the same in both spaces.

This kind of ambiguity is one of the reasons for introducing the concept of *relative topology*: we recall that a subset  $A$  of  $\mathbb{R}^n$  can be equipped with the topology relative to  $A$ , by defining its “closed balls”  $B(x, \delta) \cap A$ , for  $x \in A$ ; then  $A$  becomes a topological space in its own. In convex analysis, the topology of  $\mathbb{R}^n$  is of moderate interest: the topologies relative to *affine manifolds* turn out to be much richer.

**Definition 2.1.1** The *relative interior*  $\text{ri } C$  (or  $\text{relint } C$ ) of a convex set  $C \subset \mathbb{R}^n$  is the interior of  $C$  for the topology relative to the affine hull of  $C$ . In other words:  $x \in \text{ri } C$  if and only if

$$x \in \text{aff } C \quad \text{and} \quad \exists \delta > 0 \text{ such that } (\text{aff } C) \cap B(x, \delta) \subset C.$$

The *dimension* of a convex set  $C$  is the dimension of its affine hull, that is to say the dimension of the subspace parallel to  $\text{aff } C$ .  $\square$

Thus, the wording “relative” implicitly means by convention “relative to the affine hull”. Of course, note that  $\text{ri } C \subset C$ . All along this section, and also later in Theorem V.2.2.3, we will see that  $\text{aff } C$  is the relevant working topological space. Already now, observe that our sheet of paper above can be moved ad libitum in  $\mathbb{R}^3$  (but not folded: it would become nonconvex); its affine hull and relative interior move with it, but are otherwise unaltered. Indeed, the relative topological properties of  $C$  are the properties of convex sets in  $\mathbb{R}^k$ , where  $k$  is the dimension of  $C$  or  $\text{aff } C$ . Table 2.1.1 gives some examples.

**Table 2.1.1.** Various relative interiors

$C$	$\text{aff } C$	$\dim C$	$\text{ri } C$
$\{x\}$	$\{x\}$	0	$\{x\}$
$[x, x']$ $x \neq x'$	affine line generated by $x$ and $x'$	1	$]x, x'[$
$\Delta_n$	affine manifold of equation $e^\top \alpha = 1$	$n - 1$	$\{\alpha \in \Delta_n : \alpha_i > 0\}$
$B(x_0, \delta)$	$\mathbb{R}^n$	$n$	$\text{int } B(x_0, \delta)$

**Remark 2.1.2** The cluster points of a set  $C$  are in  $\text{aff } C$  (which is closed and contains  $C$ ), so the relative closure of  $C$  is just  $\text{cl } C$ : a notation  $\text{relcl } C$  would be superfluous. On the contrary, the boundary is affected, and we will speak of *relative boundary*:

$$\text{rbd } C := \text{cl } C \setminus \text{ri } C.$$

$\square$

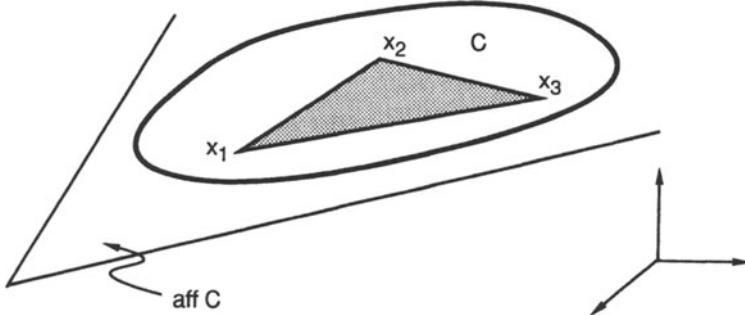
A first demonstration of the relevance of our new definition is the following:

**Theorem 2.1.3** If  $C \neq \emptyset$ , then  $\text{ri } C \neq \emptyset$ . In fact,  $\dim(\text{ri } C) = \dim C$ .

PROOF. Let  $k := 1 + \dim C$ . Since  $\text{aff } C$  has dimension  $k - 1$ ,  $C$  contains  $k$  elements affinely independent  $x_1, \dots, x_k$ . Call  $\Delta := \text{co}\{x_1, \dots, x_k\}$  the simplex that they generate; see Fig. 2.1.1;  $\text{aff } \Delta = \text{aff } C$  because  $\Delta \subset C$  and  $\dim \Delta = k - 1$ . The proof will be finished if we show that  $\Delta$  has nonempty relative interior.

Take  $\bar{x} := 1/k \sum_{i=1}^k x_i$  (the “center” of  $\Delta$ ) and describe  $\text{aff } \Delta$  by points of the form

$$\bar{x} + y = \bar{x} + \sum_{i=1}^k \alpha_i(y) x_i = \sum_{i=1}^k \left[ \frac{1}{k} + \alpha_i(y) \right] x_i,$$



**Fig. 2.1.1.** A relative interior is nonempty

where  $\alpha(y) = (\alpha_1(y), \dots, \alpha_k(y)) \in \mathbb{R}^k$  solves

$$\sum_{i=1}^k \alpha_i x_i = y, \quad \sum_{i=1}^k \alpha_i = 0.$$

Because this system has a unique solution, the mapping  $y \mapsto \alpha(y)$  is (linear and) continuous: we can find  $\delta > 0$  such that  $\|y\| \leq \delta$  implies

$$|\alpha_i(y)| \leq 1/k \text{ for } i = 1, \dots, k, \quad \text{hence } \bar{x} + y \in \Delta.$$

In other words,  $\bar{x} \in \text{ri } \Delta \subset \text{ri } C$ .

It follows in particular  $\dim \text{ri } C = \dim \Delta = \dim C$ . □

**Remark 2.1.4** We could have gone a little further in our proof, to realize that the relative interior of  $\Delta$  was

$$\left\{ \sum_{i=1}^k \alpha_i x_i : \sum_{i=1}^k \alpha_i = 1, \alpha_i > 0 \text{ for } i = 1, \dots, k \right\}.$$

Indeed, any point in the above set could have played the role of  $\bar{x}$  in the proof. Note, incidentally, that the above set is still the relative interior of  $\text{co}\{x_1, \dots, x_k\}$ , even if the  $x_i$ 's are not affinely independent. □

**Remark 2.1.5** The attention of the reader is drawn to a detail in the proof of Theorem 2.1.3:  $\Delta \subset C$  implied  $\text{ri } \Delta \subset \text{ri } C$  because  $\Delta$  and  $C$  had the same affine hull, hence the same relative topology. Taking the relative interior is not a monotone operation, though: in  $\mathbb{R}$ ,  $\{0\} \subset [0, 1]$  but  $\{0\} = \text{ri}\{0\}$  is not contained in the relative interior  $[0, 1[$  of  $[0, 1]$ . □

We now turn to a *very useful* technical result; it refines the intermediate result in the proof of Proposition 1.2.7, illustrated by Fig. 1.2.3: when moving from a point in  $\text{ri } C$  straight to a point of  $\text{cl } C$ , we stay inside  $\text{ri } C$ .

**Lemma 2.1.6** *Let  $x \in \text{cl } C$  and  $x' \in \text{ri } C$ . Then the half-open segment*

$$]x, x'] = \{\alpha x + (1 - \alpha)x' : 0 \leq \alpha < 1\}$$

*is contained in  $\text{ri } C$ .*

PROOF. Take  $x'' = \alpha x + (1 - \alpha)x'$ , with  $1 > \alpha \geq 0$ . To avoid writing “ $\cap \text{aff } C$ ” every time, we assume without loss of generality that  $\text{aff } C = \mathbb{R}^n$ .

Since  $x \in \text{cl } C$ , for all  $\varepsilon > 0$ ,  $x \in C + B(0, \varepsilon)$  and we can write

$$\begin{aligned} B(x'', \varepsilon) &= \alpha x + (1 - \alpha)x' + B(0, \varepsilon) \\ &\subset \alpha C + (1 - \alpha)x' + (1 + \alpha)B(0, \varepsilon) \\ &= \alpha C + (1 - \alpha)\{x' + B(0, \frac{1+\alpha}{1-\alpha}\varepsilon)\}. \end{aligned}$$

Since  $x' \in \text{int } C$ , we can choose  $\varepsilon$  so small that  $x' + B(0, \frac{1+\alpha}{1-\alpha}\varepsilon) \subset C$ . Then we have

$$B(x'', \varepsilon) \subset \alpha C + (1 - \alpha)C = C$$

(where the last equality is just the definition of a convex set).  $\square$

**Remark 2.1.7** We mention an interesting consequence of this result: a half-line issued from  $x' \in \text{ri } C$  cannot cut the boundary of  $C$  in more than one point; hence, a line meeting  $\text{ri } C$  cannot cut  $\text{cl } C$  in more than two points: the relative boundary of a convex set is thus a fairly regular object, looking like an “onion skin” (see Fig. 2.1.2).  $\square$

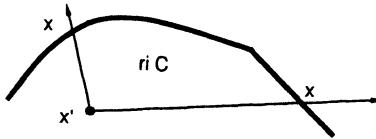


Fig. 2.1.2. The relative boundary of a convex set

Note in particular that  $[x, x'] \subset \text{ri } C$  whenever  $x$  and  $x'$  are in  $\text{ri } C$ , which confirms that  $\text{ri } C$  is convex (cf. Proposition 1.2.7). Actually,  $\text{ri } C$ ,  $C$  and  $\text{cl } C$  are three convex sets very close together: they are *not distinguished* by the operations “aff”, “ri” and “cl”.

**Proposition 2.1.8** *The three convex sets  $\text{ri } C$ ,  $C$  and  $\text{cl } C$  have the same affine hull (and hence the same dimension), the same relative interior and the same closure (and hence the same relative boundary).*

PROOF. The case of the affine hull was already seen in Theorem 2.1.3. For the others, the key result is Lemma 2.1.6 (as well as for most other properties involving closures and relative interiors). We illustrate it by restricting our proof to one of the properties, say:  $\text{ri } C$  and  $C$  have the same closure.

Thus, we have to prove that  $\text{cl } C \subset \text{cl}(\text{ri } C)$ . Let  $x \in \text{cl } C$  and take  $x' \in \text{ri } C$  (it is possible by virtue of Theorem 2.1.3). Because  $]x, x'] \subset \text{ri } C$  (Lemma 2.1.6), we do have that  $x$  is a limit of points in  $\text{ri } C$  (and even a “radial” limit); hence  $x$  is in the closure of  $\text{ri } C$ .  $\square$

**Remark 2.1.9** This result gives one more argument in favour of our relative topology: if we take a closed convex set  $C$ , open it (for the topology of  $\text{aff } C$ ), and close the result, we obtain  $C$  again – a very relevant topological property.

Among the consequences of Proposition 2.1.8, we mention the following:

- $C$  and  $\text{cl } C$  have the same interior – hence the same boundary: in fact, either both are empty (when  $\dim C = \dim \text{cl } C < n$ ), or they coincide because the interior equals the relative interior.
- If  $C_1$  and  $C_2$  are two convex sets having the same closure, then they generate the same affine manifold and have the same relative interior. This happens exactly when we have the following “sandwich” relation

$$\text{ri } C_1 \subset C_2 \subset \text{cl } C_1.$$

□

Our relative topology fits rather well with the convexity-preserving operations presented in §1.2. Our first result in these lines is of paramount importance in convex analysis and optimization.

**Proposition 2.1.10** *Let the two convex sets  $C_1$  and  $C_2$  satisfy  $\text{ri } C_1 \cap \text{ri } C_2 \neq \emptyset$ . Then*

$$\text{ri}(C_1 \cap C_2) = \text{ri } C_1 \cap \text{ri } C_2 \quad (2.1.1)$$

$$\text{cl}(C_1 \cap C_2) = \text{cl } C_1 \cap \text{cl } C_2. \quad (2.1.2)$$

PROOF. First we show that  $\text{cl } C_1 \cap \text{cl } C_2 \subset \text{cl}(C_1 \cap C_2)$  (the converse inclusion is always true). Given  $x \in \text{cl } C_1 \cap \text{cl } C_2$ , we pick  $x'$  in the nonempty  $\text{ri } C_1 \cap \text{ri } C_2$ . From Lemma 2.1.6 applied to  $C_1$  and to  $C_2$ ,

$$]x, x'] \subset \text{ri } C_1 \cap \text{ri } C_2.$$

Taking the closure of both sides, we conclude

$$x \in \text{cl}(\text{ri } C_1 \cap \text{ri } C_2) \subset \text{cl}(C_1 \cap C_2),$$

which proves (2.1.2) because  $x$  was arbitrary; the above inclusion is actually an equality.

Now, we have just seen that the two convex sets  $\text{ri } C_1 \cap \text{ri } C_2$  and  $C_1 \cap C_2$  have the same closure. According to Remark 2.1.9, they have the same relative interior:

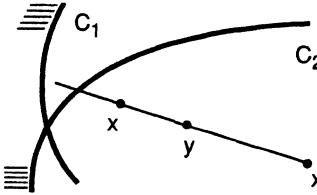
$$\text{ri}(C_1 \cap C_2) = \text{ri}(\text{ri } C_1 \cap \text{ri } C_2) \subset \text{ri } C_1 \cap \text{ri } C_2.$$

It remains to prove the converse inclusion, so let  $y \in \text{ri } C_1 \cap \text{ri } C_2$ . If we take  $x' \in C_1$  [resp.  $C_2$ ], the segment  $[x', y]$  is in  $\text{aff } C_1$  [resp.  $\text{aff } C_2$ ] and, by definition of the relative interior, this segment can be stretched beyond  $y$  and yet stay in  $C_1$  [resp.  $C_2$ ] (see Fig. 2.1.3). Take in particular  $x' \in \text{ri}(C_1 \cap C_2)$ ,  $x' \neq y$  (if such an  $x'$  does not exist, we are done). The above stretching singles out an  $x \in C_1 \cap C_2$  such that  $y \in ]x, x':[$ :

$$y = \alpha x + (1 - \alpha)x' \quad \text{for some } \alpha \in ]0, 1[.$$

Then Lemma 2.1.6 applied to  $C_1 \cap C_2$  tells us that  $y \in \text{ri}(C_1 \cap C_2)$ .

□



**Fig. 2.1.3.** The stretching mechanism

Observe that, if we intersect infinitely many convex sets – instead of two, or a finite number –, the proof of (2.1.2) still works, but certainly not the proof of (2.1.1): the stretching possibility is killed. The condition that the relative interiors have a nonempty intersection is very important and will be encountered many times in the sequel; it is essential for both (2.1.1) and (2.1.2) (use the same counter-example as in Remark 2.1.5). Incidentally, it gives another sufficient condition for the monotonicity of the ri-operation (use (2.1.1) with  $C_1 \subset C_2$ ).

We restrict our next statements to the case of the relative interior. Lemma 2.1.6 and Proposition 2.1.8 help in carrying them over to the closure operation.

**Proposition 2.1.11** *For  $i = 1, \dots, k$ , let  $C_i \subset \mathbb{R}^{n_i}$  be convex sets. Then*

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

PROOF. It suffices to apply Definition 2.1.1 alone, observing that

$$\text{aff}(C_1 \times \cdots \times C_k) = (\text{aff } C_1) \times \cdots \times (\text{aff } C_k).$$

□

**Proposition 2.1.12** *Let  $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$  be an affine mapping and  $C$  a convex set of  $\mathbb{R}^n$ . Then*

$$\text{ri}[A(C)] = A(\text{ri } C). \quad (2.1.3)$$

*If  $D$  is a convex set of  $\mathbb{R}^m$  satisfying  $\bar{A}(\text{ri } D) \neq \emptyset$ , then*

$$\text{ri}\left[\bar{A}(D)\right] = \bar{A}(\text{ri } D). \quad (2.1.4)$$

PROOF. First, note that the continuity of  $A$  implies  $A(\text{cl } S) \subset \text{cl}[A(S)]$  for any  $S \subset \mathbb{R}^n$ . Apply this result to  $\text{ri } C$ , whose closure is  $\text{cl } C$  (Proposition 2.1.8), and use the monotonicity of the closure operation:

$$A(C) \subset A(\text{cl } C) = A[\text{cl}(\text{ri } C)] \subset \text{cl}[A(\text{ri } C)] \subset \text{cl}[A(C)];$$

the closed set  $\text{cl}[A(\text{ri } C)]$  is therefore  $\text{cl}[A(C)]$ . Because  $A(\text{ri } C)$  and  $A(C)$  have the same closure, they have the same relative interior (Remark 2.1.9):

$$\text{ri } A(C) = \text{ri}[A(\text{ri } C)] \subset A(\text{ri } C).$$

To prove the converse inclusion, let  $w = A(y) \in A(\text{ri } C)$ , with  $y \in \text{ri } C$ . We choose  $z' = A(x') \in \text{ri } A(C)$ , with  $x' \in C$  (we assume  $z' \neq w$ , hence  $x' \neq y$ ).

Using in  $C$  the same stretching mechanism as in Fig. 2.1.3, we single out  $x \in C$  such that  $y \in ]x, x'[,$  to which corresponds  $z = A(x) \in A(C).$  By affinity,  $A(y) \in ]A(x), A(x')[ = ]z, z'[.$  Thus,  $z$  and  $z'$  fulfil the conditions of Lemma 2.1.6 applied to the convex set  $A(C): w \in \text{ri } A(C),$  and (2.1.3) is proved.

The proof of (2.1.4) uses the same technique.  $\square$

As an illustration of the last two results, we see that the relative interior of  $\alpha_1 C_1 + \alpha_2 C_2$  is  $\alpha_1 \text{ri } C_1 + \alpha_2 \text{ri } C_2.$  If we take in particular  $\alpha_1 = -\alpha_2 = 1,$  we obtain the following theorem:

$$0 \in \text{ri}(C_1 - C_2) \iff (\text{ri } C_1) \cap (\text{ri } C_2) \neq \emptyset, \quad (2.1.5)$$

which gives one more equivalent form for the condition in Proposition 2.1.10. We will come again to this property on several occasions.

## 2.2 The Asymptotic Cone

Let  $x$  be a point in a closed convex cone  $K.$  Draw a picture to see that, for all  $d \in K,$  the half-line  $x + \mathbb{R}^+ d$  is contained in  $K: x + td \in K$  for all  $t > 0.$  Conversely, if  $x + \mathbb{R}^+ d \subset K,$  i.e. if

$$d \in \frac{K - x}{t} = K - \left\{ \frac{1}{t}x \right\} \quad \text{for all } t > 0,$$

then ( $K$  is closed),  $d \in K.$  In words, a closed convex cone is also the set of directions along which one can go straight to infinity. We now generalize this concept to non-conical sets.

In this section,  $C$  will always be a nonempty *closed convex* set. For  $x \in C,$  let

$$C_\infty(x) := \{d \in \mathbb{R}^n : x + td \in C \text{ for all } t > 0\}. \quad (2.2.1)$$

Despite the appearances,  $C_\infty(x)$  depends only on the behaviour of  $C$  “at infinity”: in fact,  $x + td \in C$  implies that  $x + \tau d \in C$  for all  $\tau \in [0, t]$  ( $C$  is convex). Thus,  $C_\infty(x)$  is just the set of directions from which one can go straight from  $x$  to infinity, while staying in  $C.$  Another formulation is:

$$C_\infty(x) = \bigcap_{t>0} \frac{C - x}{t}, \quad (2.2.2)$$

which clearly shows that  $C_\infty(x)$  is a *closed convex cone*, which of course contains 0. The following property is fundamental.

**Proposition 2.2.1** *The closed convex cone  $C_\infty(x)$  does not depend on  $x \in C.$*

**PROOF.** See Theorem I.2.3.1 and the pantographic Figure I.2.3.1. Take two different points  $x_1$  and  $x_2$  in  $C;$  it suffices to prove one inclusion, say  $C_\infty(x_1) \subset C_\infty(x_2).$  Let  $d \in C_\infty(x_1)$  and  $t > 0,$  we have to prove  $x_2 + td \in C.$  With  $\varepsilon \in ]0, 1[,$  consider the point

$$\bar{x}_\varepsilon := x_1 + td + (1 - \varepsilon)(x_2 - x_1).$$

Writing it as

$$\bar{x}_\varepsilon = \varepsilon \left( x_1 + \frac{t}{\varepsilon} d \right) + (1 - \varepsilon) x_2,$$

we see that  $\bar{x}_\varepsilon \in C$  (use the definitions of  $C_\infty(x_1)$  and of a convex set). On the other hand,

$$x_2 + td = \lim_{\varepsilon \downarrow 0} \bar{x}_\varepsilon \in \text{cl } C = C.$$
□

It follows that the notation  $C_\infty$  is more appropriate:

**Definition 2.2.2** The *asymptotic cone*, or recession cone of the closed convex set  $C$  is the closed convex cone  $C_\infty$  defined by (2.2.1) or (2.2.2), in which Proposition 2.2.1 is exploited. □

Figure 2.2.1 gives three examples in  $\mathbb{R}^2$ . As for the asymptotic cone of Example 1.2.2, it is the set  $\{d \in \mathbb{R}^n : Ad \leq 0\}$ .

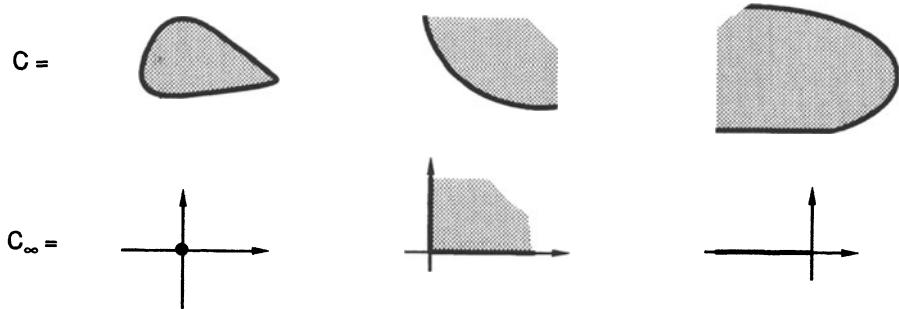


Fig. 2.2.1. Some asymptotic cones

**Proposition 2.2.3** A closed convex set  $C$  is compact if and only if  $C_\infty = \{0\}$ .

PROOF. If  $C$  is bounded, it is clear that  $C_\infty$  cannot contain any nonzero direction. Conversely, let  $\{x_k\} \subset C$  be such that  $\|x_k\| \rightarrow +\infty$  (we assume  $x_k \neq 0$ ). The sequence  $\{d_k := x_k/\|x_k\|\}$  is bounded, extract a convergent subsequence:  $d = \lim_{k \in K} d_k$  with  $K \subset \mathbb{N}$  ( $\|d\| = 1$ ). Now, given  $x \in C$  and  $t > 0$ , take  $k$  so large that  $\|x_k\| \geq t$ . Then, we see that

$$x + td = \lim_{k \in K} \left[ \left( 1 - \frac{t}{\|x_k\|} \right) x + \frac{t}{\|x_k\|} x_k \right]$$

is in the closed convex set  $C$ , hence  $d \in C_\infty$ . □

Another easy-to-see relationship is

$$C_\infty = \{d \in \mathbb{R}^n : d + C \subset C\} = C \pm C,$$

where the star-difference is that of Example 1.2.6. It follows that  $C_\infty$  can be viewed as the maximal  $X \subset \mathbb{R}^n$  (in the sense of the  $\subset$ -relation) solving the set-valued equation

$$X + C = C \quad [\text{or equivalently } X + C \subset C],$$

whose solution is  $C$  if  $C$  is a cone (see the introduction to this Section 2.2).

**Remark 2.2.4** Consider the closed convex sets  $(C - x)/t$ , indexed by  $t > 0$ . They form a nested decreasing family: for  $t_1 < t_2$  and  $y$  arbitrary in  $C$ ,

$$\frac{y - x}{t_2} = \frac{y' - x}{t_1} \quad \text{where} \quad y' := \frac{t_2 - t_1}{t_2}x + \frac{t_1}{t_2}y \in C.$$

Thus, we can write (see §A.5 for the set-limit appearing below)

$$C_\infty = \bigcap_{t>0} \frac{C - x}{t} = \lim_{t \rightarrow +\infty} \frac{C - x}{t}, \quad (2.2.3)$$

which interprets  $C_\infty$  as a limit of set-valued difference quotients, but with the denominator tending to  $\infty$ , instead of the usual 0. This will be seen again later in §5.2.  $\square$

In contrast to the relative interior, the concept of asymptotic cone does not always fit well with usual convexity-preserving operations. We just mention some properties which result directly from the definition of  $C_\infty$ .

### Proposition 2.2.5

– If  $\{C_j\}_{j \in J}$  is a family of closed convex sets having a point in common, then

$$(\cap_{j \in J} C_j)_\infty = \cap_{j \in J} (C_j)_\infty.$$

– If, for  $j = 1, \dots, m$ ,  $C_j$  are closed convex sets in  $\mathbb{R}^{n_j}$ , then

$$(C_1 \times \dots \times C_m)_\infty = (C_1)_\infty \times \dots \times (C_m)_\infty.$$

– Let  $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$  be an affine mapping. If  $C$  is closed convex in  $\mathbb{R}^n$  and  $A(C)$  is closed, then

$$A(C_\infty) \subset [A(C)]_\infty.$$

– If  $D$  is closed convex in  $\mathbb{R}^m$  with nonempty inverse image, then

$$\left[ A^{-1}(D) \right]_\infty = A^{-1}(D_\infty).$$

$\square$

Needless to say, convexity does not help to ensure that the image of a closed set under a continuous mapping is closed: take  $A(\xi, \eta) = \xi$  (linear) and  $C = \{(\xi, \eta) : \eta \geq 1/\xi > 0\}$ .

## 2.3 Extreme Points

In this section,  $C$  is a nonempty convex set of  $\mathbb{R}^n$  and there would be no loss of generality in assuming that it is closed. The reader may make this assumption if he finds it helpful in mastering faster the definitions and properties below; the same remark holds for §2.4.

**Definition 2.3.1** We say that  $x \in C$  is an *extreme point* of  $C$  if there are no two different points  $x_1$  and  $x_2$  in  $C$  such that  $x = 1/2(x_1 + x_2)$ .  $\square$

Some other ways of expressing the same thing are:

- $x = \alpha x_1 + (1 - \alpha)x_2$  is impossible whenever  $x_1$  and  $x_2$  are two distinct points of  $C$  and  $\alpha \in ]0, 1[$ : indeed, convexity of  $C$  implies that  $x_1$  and  $x_2$  in the definition can be replaced by two other points in the segment  $[x_1, x_2]$ ; this amounts to replacing the number  $1/2$  by some other  $\alpha \in ]0, 1[$ . In short:

$$\begin{aligned} x \text{ is an extreme point of } C \text{ if and only if} \\ [x = \alpha x_1 + (1 - \alpha)x_2, x_i \in C, 0 < \alpha < 1] \implies x = x_1 = x_2. \end{aligned}$$

- There is no convex combination  $x = \sum_{i=1}^k \alpha_i x_i$  other than  $x_1 = \dots = x_k [= x]$ .
- The set  $C \setminus \{x\}$  is still convex.

### Examples 2.3.2

- Let  $C$  be the unit ball  $B(0, 1)$ . Multiply by  $1/2$  the relation

$$\frac{1}{2}\|x_1 + x_2\|^2 = \|x_1\|^2 + \|x_2\|^2 - \frac{1}{2}\|x_2 - x_1\|^2 \quad (2.3.1)$$

to realize that every  $x$  of norm 1 is an extreme point of  $B(0, 1)$ . Likewise, if  $Q : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is a positive definite symmetric linear operator, any  $x$  with  $\langle Qx, x \rangle = 1$  is an extreme point of the convex set

$$\{x \in \mathbb{R}^n : \langle Qx, x \rangle \leqslant 1\}.$$

On the other hand, if  $\langle Q \cdot, \cdot \rangle^{1/2}$  is replaced by the  $\ell_1$ -norm, the corresponding unit ball has finitely many extreme points.

- If  $C$  is a convex cone, a nonzero  $x \in C$  has no chance of being an extreme point.
- An affine manifold, a half-space have no extreme points.  $\square$

The *set of extreme points* of  $C$  will be denoted by  $\text{ext } C$ . We mention here that it is a closed set when  $n \leqslant 2$ ; but in general,  $\text{ext } C$  has no particular topological or linear properties. Along the lines of the above examples, there is at least one case where there exist extreme points:

**Proposition 2.3.3** *If  $C$  is compact, then  $\text{ext } C \neq \emptyset$ .*

PROOF. Because  $C$  is compact, there is  $\bar{x} \in C$  maximizing the continuous function  $x \mapsto \|x\|^2$ . We claim that  $\bar{x}$  is extremal. In fact, suppose that there are  $x_1$  and  $x_2$  in  $C$  with  $\bar{x} = 1/2(x_1 + x_2)$ . Then, with  $x_1 \neq x_2$  and using (2.3.1), we obtain the contradiction

$$\|\bar{x}\|^2 = \left\| \frac{1}{2}(x_1 + x_2) \right\|^2 < \frac{1}{2}(\|x_1\|^2 + \|x_2\|^2) \leqslant \frac{1}{2}(\|\bar{x}\|^2 + \|\bar{x}\|^2) = \|\bar{x}\|^2. \quad \square$$

The definitions clearly imply that any extreme point of  $C$  is on its boundary, and even on its relative boundary. The essential result on extreme points is the following, which we will prove later in §4.2(c).

**Theorem 2.3.4 (H. Minkowski)** *Let  $C$  be compact, convex in  $\mathbb{R}^n$ . Then  $C$  is the convex hull of its extreme points:  $C = \text{co}(\text{ext } C)$ .*  $\square$

Combined with Carathéodory's Theorem 1.3.6, this result establishes that, if  $\dim C = k$ , then any element of  $C$  is a convex combination of at most  $k + 1$  extreme points of  $C$ .

**Example 2.3.5** Take  $C = \text{co}\{x_1, \dots, x_m\}$ . All the extreme points of  $C$  are present in the list  $x_1, \dots, x_m$ ; but of course, the  $x_i$ 's are not all necessarily extremal. Let  $\mu \leq m$  be the number of extreme points of  $C$ , suppose to simplify that these are  $x_1, \dots, x_\mu$ . Then  $C = \text{co}\{x_1, \dots, x_\mu\}$  and this representation is *minimal*, in the sense that removing one of the generators  $x_1, \dots, x_\mu$  effectively changes  $C$ . The case  $\mu = n + 1$  corresponds to a simplex in  $\mathbb{R}^n$ . If  $\mu > n + 1$ , then for any  $x \in C$ , there is a representation  $x = \sum_{i=1}^{\mu} \alpha_i x_i$  in which at least  $\mu - n - 1$  among the  $\alpha'_i$ 's are zero.  $\square$

A higher-dimensional generalization of extreme points can be defined. Consider again Definition 2.3.1, and replace “the point  $x \in C$ ” by “the convex subset  $F \subset C$ ”. Our definition is then generalized as follows: the convex subset  $F \subset C$  is *extremal* if there are no two points  $x_1$  and  $x_2$  in  $C \setminus F$  such that  $\frac{1}{2}(x_1 + x_2) \in F$ .

Once again, the number  $\frac{1}{2}$  has nothing special and can be replaced by any other  $\alpha \in ]0, 1[$ . The above statement can be rephrased in reversed logic as: if  $x_1$  and  $x_2$  in  $C$  are such that  $\alpha x_1 + (1 - \alpha)x_2 \in F$  for some  $\alpha \in ]0, 1[$ , then  $x_1$  and  $x_2$  are in  $F$  as well. Convexity of  $F$  then implies that the whole segment  $[x_1, x_2]$  is in  $F$ , and we end up with the traditional definition:

**Definition 2.3.6** A nonempty convex subset  $F \subset C$  is a *face* of  $C$  if it satisfies the following property: every segment of  $C$ , having in its relative interior an element of  $F$ , is entirely contained in  $F$ . In other words,

$$\left. \begin{array}{l} (x_1, x_2) \in C \times C \quad \text{and} \\ \exists \alpha \in ]0, 1[ : \alpha x_1 + (1 - \alpha)x_2 \in F \end{array} \right\} \implies [x_1, x_2] \subset F. \quad (2.3.2) \quad \square$$

Being convex, a face has its own affine hull, closure, relative interior and dimension. By construction, extreme points appear as faces that are singletons, i.e. 0-dimensional faces:

$$x \in \text{ext } C \iff \{x\} \text{ is a face of } C.$$

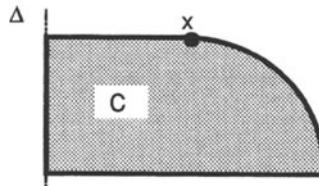
One-dimensional faces, i.e. segments that are faces of  $C$ , are called *edges* of  $C$ ; and so on until  $(k - 1)$ -dimensional faces (where  $k = \dim C$ ), called *facets* ... and the only  $k$ -dimensional face of  $C$ , which is  $C$  itself.

A useful property is the “transmission of extremality”: if  $x \in C' \subset C$  is an extreme point of  $C$ , then it is a fortiori an extreme point of the smaller set  $C'$ . When  $C'$  is a face of  $C$ , the converse is also true:

**Proposition 2.3.7** *Let  $F$  be a face of  $C$ . Then any extreme point of  $F$  is an extreme point of  $C$ .*

PROOF. Take  $x \in F \subset C$  and assume that  $x$  is not an extreme point of  $C$ : there are different  $x_1, x_2$  in  $C$  and  $\alpha \in ]0, 1[$  such that  $x = \alpha x_1 + (1 - \alpha)x_2 \in F$ . From the very definition (2.3.2) of a face, this implies that  $x_1$  and  $x_2$  are in  $F$ :  $x$  cannot be an extreme point of  $F$ .  $\square$

This property can be generalized to: if  $F'$  is a face of  $F$ , which is itself a face of  $C$ , then  $F'$  is a face of  $C$ . We mention also: the relative interiors of the faces of  $C$  form a partition of  $C$ . Examine Example 2.3.5 to visualize its faces, their relative interiors and the above partition. The  $C$  of Fig. 2.3.1, with its extreme point  $x$ , gives a less trivial situation; make a three-dimensional convex set by rotating  $C$  around the axis  $\Delta$ : we obtain a set with no one-dimensional face.



**Fig. 2.3.1.** A special extreme point

Faces (other than extreme points) are not too important in convex analysis and optimization – and this is fortunate: after all, Definition 2.3.6 is rather tricky. A much more useful concept is that of exposed faces, the subject of the next section.

## 2.4 Exposed Faces

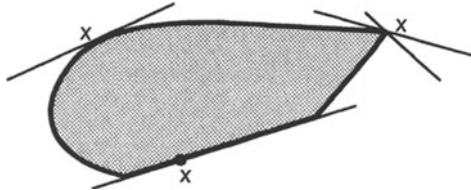
The rationale for extreme points is an inner construction of convex sets, as is particularly illustrated by Theorem 2.3.4 and Example 2.3.5. We mentioned in the important Remark 1.4.4 that a convex set could also be constructed externally, by taking intersections of convex sets containing it (see Proposition 1.3.4: if  $S$  is convex, then  $S = \text{co } S$ ). To prepare a deeper analysis, coming in §4.2(b) and §5.2, we need the following fundamental definition, based on Example 1.1.2.

**Definition 2.4.1 (Supporting Hyperplane)** An affine hyperplane  $H_{s,r}$  is said to *support* the set  $C$  when  $C$  is entirely contained in one of the two closed half-spaces delimited by  $H_{s,r}$ : say

$$\langle s, y \rangle \leq r \quad \text{for all } y \in C. \quad (2.4.1)$$

It is said to support  $C$  at  $x \in C$  when, in addition,  $x \in H_{s,r}$ : (2.4.1) holds, as well as  $\langle s, x \rangle = r$ .  $\square$

See Fig. 2.4.1 for an illustration. Up to now, it is only a formal definition; existence of some supporting hyperplane will be established later in §4.2(a). Naturally, the inequality-sign could be reversed in (2.4.1):  $H_{s,r}$  supports  $C$  when  $H_{-s,-r}$  supports  $C$ . Note also that if  $x \in C$  has a hyperplane supporting  $C$ , then  $x \in \text{bd } C$ .



**Fig. 2.4.1.** Supporting hyperplanes at various points

**Definition 2.4.2** The set  $F \subset C$  is an *exposed face* of  $C$  if there is a supporting hyperplane  $H_{s,r}$  of  $C$  such that  $F = C \cap H_{s,r}$ .

An *exposed point*, or *vertex*, is a 0-dimensional exposed face, i.e. a point  $x \in C$  at which there is a supporting hyperplane  $H_{s,r}$  of  $C$  such that  $H_{s,r} \cap C$  reduces to  $\{x\}$ .  $\square$

See Fig. 2.4.1 again. A supporting hyperplane  $H_{s,r}$  may or may not touch  $C$ . If it does, the contact-set is an exposed face. If it does at a singleton, this singleton is called an exposed point. As an intersection of convex sets, an exposed face is convex. The next result justifies the wording.

**Proposition 2.4.3** *An exposed face is a face.*

PROOF. Let  $F$  be an exposed face, with its associated support  $H_{s,r}$ . Take  $x_1$  and  $x_2$  in  $C$ :

$$\langle s, x_i \rangle \leq r \quad \text{for } i = 1, 2; \tag{2.4.2}$$

take also  $\alpha \in ]0, 1[$  such that  $\alpha x_1 + (1 - \alpha)x_2 \in F \subset H_{s,r}$ :

$$\langle s, \alpha x_1 + (1 - \alpha)x_2 \rangle = r.$$

Suppose that one of the relations (2.4.2) holds as strict inequality. By convex combination, we obtain ( $0 < \alpha < 1!$ )

$$\langle s, \alpha x_1 + (1 - \alpha)x_2 \rangle < r,$$

a contradiction.  $\square$

The simple technique used in the above proof appears often in convex analysis: if a convex combination, with positive coefficients, of inequalities holds as an equality, then so does each individual inequality.

**Remark 2.4.4** Comparing with Proposition 2.3.7, we see that the property of transmission of extremality applies to exposed faces as well: if  $x$  is an extreme point of the exposed face  $F \subset C$ , then  $x \in \text{ext } C$ .  $\square$

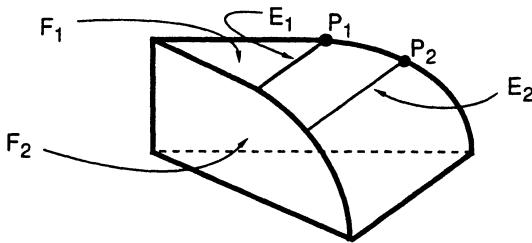
One could believe (for example from Fig. 2.4.1) that the converse to Proposition 2.4.3 is true. Figure 2.3.1 immediately shows that this intuition is false: the extreme point  $x$  is not exposed. Exposed faces form therefore a proper subset of faces. The difference is slight,

however: a result of S. Straszewicz (1935) establishes that any extreme point of a closed convex set  $C$  is a limit of exposed points in  $C$ . In other words,

$$\exp C \subset \text{ext } C \subset \text{cl}(\exp C)$$

if  $\exp C$  denotes the set of exposed points in  $C$ . Comparing with Minkowski's result 2.3.4, we see that  $C = \overline{\text{co}}(\exp C)$  for  $C$  convex and compact. We also mention that a facet is automatically exposed (the reason is that  $n - 1$ , the dimension of a facet, is also the dimension of the hyperplane involved when exposing faces).

Enrich Fig. 2.3.1 as follows: take  $C'$ , obtained from  $C$  by a rotation of  $30^\circ$  around  $\Delta$ ; then consider the convex hull of  $C \cup C'$ , displayed in Fig. 2.4.2. The point  $P_1$  is an extreme point but not a vertex;  $P_2$  is a vertex. The edge  $E_1$  is not exposed;  $E_2$  is an exposed edge. As for the faces  $F_1$  and  $F_2$ , they are exposed because they are facets.



**Fig. 2.4.2.** Faces and exposed faces

**Remark 2.4.5 (Direction Exposing a Face)** Let  $F$  be an exposed face, and  $H_{s,r}$  its associated supporting hyperplane. It results immediately from the definitions that

$$\langle s, y \rangle \leq \langle s, x \rangle \quad \text{for all } y \in C \text{ and all } x \in F.$$

Another definition of an exposed face can therefore be proposed, as the set of maximizers over  $C$  of some linear form:  $F$  is an exposed face of  $C$  when there is a nonzero  $s \in \mathbb{R}^n$  such that

$$F = \left\{ x \in C : \langle s, x \rangle = \sup_{y \in C} \langle s, y \rangle \right\}. \quad (2.4.3)$$

A relevant notation is thus  $F_C(s)$  to designate the exposed face of  $C$  associated with  $s \in \mathbb{R}^n$ ; it can also be called the face of  $C$  exposed by  $s$ . For a unified notation, we will consider  $C$  itself as exposed by 0:  $C = F_C(0)$ .  $\square$

Beware that a given  $s$  may define no supporting hyperplane at all. Even if it does, it may expose no face (the supremum in (2.4.3) may be not attained). The following result is almost trivial, but very useful: it is “equivalent” to extremize a linear form on a compact set or on its convex hull.

**Proposition 2.4.6** *Let  $C$  be convex and compact. For  $s \in \mathbb{R}^n$ , there holds*

$$\max_{x \in C} \langle s, x \rangle = \max_{x \in \text{ext } C} \langle s, x \rangle.$$

Furthermore, the solution-set of the first problem is the convex hull of the solution-set of the second:

$$\operatorname{Argmax}_{x \in C} \langle s, x \rangle = \text{co} \left\{ \operatorname{Argmax}_{x \in \text{ext } C} \langle s, x \rangle \right\}.$$

PROOF. Because  $C$  is compact,  $\langle s, \cdot \rangle$  attains its maximum on  $F_C(s)$ . The latter set is convex and compact, and as such is the convex hull of its extreme points (Minkowski's Theorem 2.3.4); these extreme points are also extreme in  $C$  (Proposition 2.3.7 and Remark 2.4.4).  $\square$

### 3 Projection onto Closed Convex Sets

#### 3.1 The Projection Operator

Denote by  $p_V$  the (orthogonal) projection onto a subspace  $V \subset \mathbb{R}^n$ . The main properties of the operator  $x \mapsto p_V(x)$  are to be linear, symmetric, positive semi-definite, idempotent ( $p_V \circ p_V = p_V$ ), nonexpansive ( $\|p_V(x)\| \leq \|x\|$  for all  $x$ ); also, it defines a canonical decomposition of  $\mathbb{R}^n$  via  $x = p_V(x) + p_{V^\perp}(x)$ . We will generalize this operator to the case where  $V$  is merely convex.

In what follows,  $C$  is a *nonempty closed convex* set of  $\mathbb{R}^n$ . For fixed  $x \in \mathbb{R}^n$ , we consider the following problem:

$$\inf \left\{ \frac{1}{2} \|y - x\|^2 : y \in C \right\}, \quad (3.1.1)$$

i.e. we are interested in those points (if any) of  $C$  that are closest to  $x$  for the Euclidean distance. Let  $f_x : \mathbb{R}^n \rightarrow \mathbb{R}$  be the function which, to  $y \in \mathbb{R}^n$ , associates

$$f_x(y) := \frac{1}{2} \|y - x\|^2. \quad (3.1.2)$$

For  $c \in C$ , take the sublevel-set  $S := \{y \in \mathbb{R}^n : f_x(y) \leq f_x(c)\}$ . Then (3.1.1) is clearly equivalent to

$$\inf \{f_x(y) : y \in C \cap S\},$$

which has a solution since  $f_x$  is continuous and  $S$  – hence  $C \cap S$  – is compact. We deduce the *existence* of a closest point in  $C$  to  $x$ ; the inf in (3.1.1) is a min.

Note that convexity of  $C$  plays no role in the above existence result. Uniqueness, however, depends crucially on convexity: let  $y_1$  and  $y_2$  be two solutions to (3.1.1). Use (2.3.1) with  $x_i = y_i - x$  to obtain

$$f_x(y_0) = \frac{1}{2}[f_x(y_1) + f_x(y_2)] - \frac{1}{8} \|y_2 - y_1\|^2,$$

where  $y_0 := \frac{1}{2}(y_1 + y_2) \in C$ ; this implies *uniqueness*.

We have thus defined a *projection operator*, namely the mapping  $x \mapsto p_C(x)$  which, to each  $x \in \mathbb{R}^n$ , associates the unique solution  $p_C(x)$  of the minimization problem (3.1.1). It is possible to *characterize*  $p_C(x)$  differently, as solving a so-called *variational inequality*; and this characterization is the key to all results concerning  $p_C$ .

**Theorem 3.1.1** A point  $y_x \in C$  is the projection  $p_C(x)$  if and only if

$$\langle x - y_x, y - y_x \rangle \leq 0 \quad \text{for all } y \in C. \quad (3.1.3)$$

PROOF. Call  $y_x$  the solution of (3.1.1); take  $y$  arbitrary in  $C$ , so that  $y_x + \alpha(y - y_x) \in C$  for any  $\alpha \in ]0, 1[$ . Then we can write with the notation (3.1.2)

$$f_x(y_x) \leq f_x(y_x + \alpha(y - y_x)) = \frac{1}{2} \|y_x - x + \alpha(y - y_x)\|^2.$$

Developing the square, we obtain after simplification

$$0 \leq \alpha \langle y_x - x, y - y_x \rangle + \frac{1}{2} \alpha^2 \|y - y_x\|^2.$$

Divide by  $\alpha (> 0)$  and let  $\alpha \downarrow 0$  to obtain (3.1.3).

Conversely, suppose that  $y_x \in C$  satisfies (3.1.3). If  $y_x = x$ , then  $y_x$  certainly solves (3.1.1). If not, write for arbitrary  $y \in C$ :

$$\begin{aligned} 0 &\geq \langle x - y_x, y - y_x \rangle = \langle x - y_x, y - x + x - y_x \rangle = \\ &= \|x - y_x\|^2 + \langle x - y_x, y - x \rangle \geq \|x - y_x\|^2 - \|x - y\| \|x - y_x\|, \end{aligned}$$

where the Cauchy-Schwarz inequality is used. Divide by  $\|x - y_x\| > 0$  to see that  $y_x$  solves (3.1.1).  $\square$

Incidentally, this result proves at the same time that the variational inequality (3.1.3) has a unique solution in  $C$ . Figure 3.1.1 illustrates the following geometric interpretation: the Cauchy-Schwarz inequality defines the angle  $\theta \in [0, \pi]$  of two nonzero vectors  $u$  and  $v$  by

$$\cos \theta := \frac{\langle u, v \rangle}{\|u\| \|v\|} \in [-1, +1].$$

Then (3.1.3) expresses the fact that the angle between  $y - y_x$  and  $x - y_x$  is obtuse, for any  $y \in C$ . Writing (3.1.3) as

$$\langle x - p_C(x), y \rangle \leq \langle x - p_C(x), p_C(x) \rangle \quad \text{for all } y \in C, \quad (3.1.4)$$

we see that  $p_C(x)$  lies in the face of  $C$  exposed by  $x - p_C(x)$ .

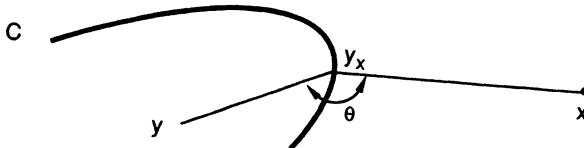


Fig. 3.1.1. The angle-characterization of a projection

**Remark 3.1.2** Suppose that  $C$  is actually an affine manifold (for example a subspace); then  $y_x - y \in C$  whenever  $y - y_x \in C$ . In this case, (3.1.3) implies that

$$\langle x - y_x, y - y_x \rangle = 0 \quad \text{for all } y \in C. \quad (3.1.5)$$

We are back with the classical characterization of the projection onto a subspace, namely that  $x - y_x \in C^\perp$  (the subspace orthogonal to  $C$ ). Passing from (3.1.3) to (3.1.5) shows once more that convex analysis is the realm of inequalities, in contrast with linear analysis.  $\square$

Some obvious properties of our projection operator are:

- the set  $\{x \in \mathbb{R}^n : p_C(x) = x\}$  of fixed points of  $p_C$  is  $C$  itself;
- from which it results that  $p_C \circ p_C = p_C$ , and also that
- $p_C$  is a linear operator if and only if  $C$  is a subspace.

More interesting is the following result:

**Proposition 3.1.3** *For all  $(x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n$ , there holds*

$$\|p_C(x_1) - p_C(x_2)\|^2 \leq \langle p_C(x_1) - p_C(x_2), x_1 - x_2 \rangle.$$

PROOF. Write (3.1.3) with  $x = x_1$ ,  $y = p_C(x_2) \in C$ :

$$\langle p_C(x_2) - p_C(x_1), x_1 - p_C(x_1) \rangle \leq 0;$$

likewise,

$$\langle p_C(x_1) - p_C(x_2), x_2 - p_C(x_2) \rangle \leq 0,$$

and conclude by addition

$$\langle p_C(x_1) - p_C(x_2), x_2 - x_1 + p_C(x_1) - p_C(x_2) \rangle \leq 0.$$

□

Two immediate consequences are worth noting. One is that

$$0 \leq \langle p_C(x_1) - p_C(x_2), x_1 - x_2 \rangle \quad \text{for all } (x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n,$$

a property expressing that the mapping  $p_C$  is, in a way, “monotone increasing”. Second, we obtain from the Cauchy-Schwarz inequality:

$$\|p_C(x_1) - p_C(x_2)\| \leq \|x_1 - x_2\|, \tag{3.1.6}$$

i.e.  $p_C$  is nonexpansive; in particular,  $\|p_C(x)\| \leq \|x\|$  whenever  $0 \in C$ . However, it is not a contraction: the best Lipschitz constant

$$L := \sup \left\{ \frac{\|p_C(x_1) - p_C(x_2)\|}{\|x_1 - x_2\|} : x_1 \neq x_2, x_1 \text{ and } x_2 \text{ out of } C \right\}$$

is equal to 1 (suppose  $C$  is a subspace!), unless more is known about the “curvature” of  $C$ .

### 3.2 Projection onto a Closed Convex Cone

As already mentioned in Example 1.1.4, convex cones are important instances of convex sets, somehow intermediate between subspaces and general convex sets. As a result, the projection operator onto a closed convex cone enjoys properties which are finer than those of §3.1, and which come closer to those of the projection onto a subspace.

**Definition 3.2.1** Let  $K$  be a convex cone, as defined in Example 1.1.4. The *polar cone* of  $K$  (called negative polar cone by some authors) is:

$$K^\circ := \{s \in \mathbb{R}^n : \langle s, x \rangle \leq 0 \text{ for all } x \in K\}. \quad \square$$

A very first observation is that the polar cone depends on the scalar product: changing  $\langle \cdot, \cdot \rangle$  changes  $K^\circ$ . One easily sees that  $K^\circ$  is a closed convex cone (use in particular continuity of the scalar product). If  $K$  is simply a subspace, then  $K^\circ$  is its orthogonal  $K^\perp$ : polarity generalizes orthogonality, remember Remark 3.1.2. Incidentally, it will be seen later in §4.2(d) that the polar of  $K^\circ$  is nothing but the closure of  $K$ . Polarity establishes a correspondence in the set of closed convex cones, which is order-reversing:

$$K' \subset K \implies (K')^\circ \supset K^\circ$$

(and the converse is true if the relation  $K^{\circ\circ} = K$  is admitted for  $K$  closed). Finally, the only possible element in  $K \cap K^\circ$  is 0.

**Examples 3.2.2** (see Fig. 3.2.1).

(a) For given  $x_1, \dots, x_m$  in  $\mathbb{R}^n$ , take the conical hull of  $m$  points  $x_1, \dots, x_m$  in  $\mathbb{R}^n$ :

$$K = \left\{ \sum_{j=1}^m \alpha_j x_j : \alpha_j \geq 0 \text{ for } j = 1, \dots, m \right\}.$$

We leave it as an exercise to check the important result:

$$K^\circ = \{s \in \mathbb{R}^n : \langle s, x_j \rangle \leq 0 \text{ for } j = 1, \dots, m\}.$$

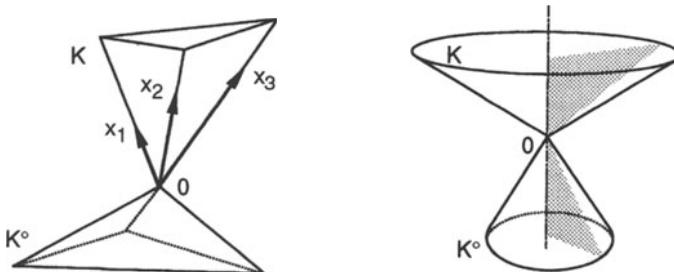


Fig. 3.2.1. Examples of polar cones

(b) As a particular case, take the usual dot-product for  $\langle \cdot, \cdot \rangle$ ,  $\mathbb{R}^n$  being equipped with the canonical basis. Then the polar of the nonnegative orthant

$$\Omega_+ := \{x = (\xi^1, \dots, \xi^n) : \xi^i \geq 0 \text{ for } i = 1, \dots, n\}$$

is the nonpositive orthant

$$\Omega_- = (\Omega_+)^{\circ} = \{s = (\sigma_1, \dots, \sigma_n) : \sigma_i \leq 0 \text{ for } i = 1, \dots, n\}.$$

Naturally, such a symmetry is purely due to the fact that the basis vectors are mutually orthogonal.

(c) Let  $K$  be a revolution cone: with  $s \in \mathbb{R}^n$  of norm 1 and  $\theta \in [0, \pi/2]$ ,

$$K_s(\theta) := \{x \in \mathbb{R}^n : \langle s, x \rangle \geq \|x\| \cos \theta\}.$$

Then  $[K_s(\theta)]^{\circ} = K_{-s}(\pi/2 - \theta)$ . □

The characterization 3.1.1 takes a special form in the present conical situation.

**Proposition 3.2.3** *Let  $K$  be a closed convex cone. Then  $y_x = p_K(x)$  if and only if*

$$y_x \in K, \quad x - y_x \in K^{\circ}, \quad \langle x - y_x, y_x \rangle = 0. \quad (3.2.1)$$

PROOF. We know from Theorem 3.1.1 that  $y_x = p_K(x)$  satisfies

$$\langle x - y_x, y - y_x \rangle \leq 0 \quad \text{for all } y \in K. \quad (3.2.2)$$

Taking  $y = \alpha y_x$ , with arbitrary  $\alpha \geq 0$ , this inequality implies

$$(\alpha - 1)\langle x - y_x, y_x \rangle \leq 0 \quad \text{for all } \alpha \geq 0.$$

Since  $\alpha - 1$  can have either sign, this implies  $\langle x - y_x, y_x \rangle = 0$  and (3.2.2) becomes

$$\langle y, x - y_x \rangle \leq 0 \quad \text{for all } y \in K, \quad \text{i.e. } x - y_x \in K^{\circ}.$$

Conversely, let  $y_x$  satisfy (3.2.1). For arbitrary  $y \in K$ , use the notation (3.1.2):

$$f_x(y) = \frac{1}{2}\|x - y_x + y_x - y\|^2 \geq f_x(y_x) + \langle x - y_x, y_x - y \rangle;$$

but (3.2.1) shows that

$$\langle x - y_x, y_x - y \rangle = -\langle x - y_x, y \rangle \geq 0,$$

hence  $f_x(y) \geq f_x(y_x)$ :  $y_x$  solves (3.1.1). □

**Remark 3.2.4** We already know from (3.1.4) that  $p_K(x)$  lies in the face of  $K$  exposed by  $x - p_K(x)$ ; but (3.2.1) tells us more: by definition of a polar cone,  $x - p_K(x)$  is also in the face of  $K^{\circ}$  exposed by  $p_K(x)$  (a symmetry confirming that  $K^{\circ\circ} = K$ ).

Take for an illustration  $K = \Omega_+$  of Example 3.2.2(b): denote by  $(\pi^1, \dots, \pi^n)$  the coordinates of  $p_K(x)$ . They are nonnegative because  $p_K(x) \in \Omega_+$ ; each term  $(\xi^i - \pi^i)\pi^i$  is nonpositive because  $x - p_K(x) \in \Omega_-$ . Because their sum is zero, each of these terms is actually zero, i.e.

$$\text{For } i = 1, \dots, n, \quad \xi^i - \pi^i = 0 \text{ or } \pi^i = 0 \text{ (or both).}$$

This property is usually called a *transversality condition*. Thus, we have:

$$\text{For each } i, \quad \text{either } \pi^i = \xi^i \text{ or } \pi^i = 0;$$

taking the nonnegativity of  $\pi$  into account, we obtain the explicit formula

$$\pi^i = \max \{0, \xi^i\} \quad \text{for } i = 1, \dots, n.$$

This implies in particular that  $\pi^i - \xi^i \geq 0$ , i.e.  $x - \pi \in \Omega_-$ . □

We list some properties which are immediate consequences of the characterization (3.2.1): for all  $x \in \mathbb{R}^n$ ,

$$\begin{aligned}\mathbf{p}_K(x) &= 0 \quad \text{if and only if } x \in K^\circ; \\ \mathbf{p}_K(\alpha x) &= \alpha \mathbf{p}_K(x) \quad \text{for all } \alpha \geq 0; \\ \mathbf{p}_K(-x) &= -\mathbf{p}_{-K}(x).\end{aligned}$$

They somehow generalize the linearity of the projection onto a subspace  $V$ . An additional property can be proved, using the obvious relation  $(-K)^\circ = -K^\circ$ :

$$\mathbf{p}_K(x) + \mathbf{p}_{K^\circ}(x) = x. \quad (3.2.3)$$

It plays the role of  $\mathbf{p}_V(x) + \mathbf{p}_{V^\perp}(x) = x$  and connotes the following *decomposition* theorem, generalizing the property  $\mathbb{R}^n = V \oplus V^\perp$ .

**Theorem 3.2.5 (J.-J. Moreau)** *Let  $K$  be a closed convex cone. For the three elements  $x, x_1$  and  $x_2$  in  $\mathbb{R}^n$ , the properties below are equivalent:*

- (i)  $x = x_1 + x_2$  with  $x_1 \in K$ ,  $x_2 \in K^\circ$  and  $\langle x_1, x_2 \rangle = 0$ ;
- (ii)  $x_1 = \mathbf{p}_K(x)$  and  $x_2 = \mathbf{p}_{K^\circ}(x)$ .

PROOF. Straightforward, from (3.2.3) and the characterization (3.2.1) of  $x_1 = \mathbf{p}_K(x)$ .  $\square$

In contrast with the decomposition in subspaces, the decomposition  $x = x_1 + x_2$ , with  $x_1 \in K$  and  $x_2 \in K^\circ$  is not unique because orthogonality is not automatic; but the decomposition (i), (ii) is *optimal* in the sense that

$$\left. \begin{array}{l} x = x_1 + x_2 \\ \text{with} \\ x_1 \in K \text{ and } x_2 \in K^\circ \end{array} \right\} \implies \left\{ \begin{array}{l} \|x_1\| \geq \|\mathbf{p}_K(x)\| \\ \text{and} \\ \|x_2\| \geq \|\mathbf{p}_{K^\circ}(x)\|. \end{array} \right.$$

## 4 Separation and Applications

### 4.1 Separation Between Convex Sets

Take two disjoint sets  $S_1$  and  $S_2$ :  $S_1 \cap S_2 = \emptyset$ . If, in addition,  $S_1$  and  $S_2$  are convex, some more can be said: a simple convex set (namely an affine hyperplane) can be squeezed between  $S_1$  and  $S_2$ . This extremely important property follows directly from those of the projection operator onto a convex set.

**Theorem 4.1.1** *Let  $C \subset \mathbb{R}^n$  be nonempty closed convex, and let  $x \notin C$ . Then there exists  $s \in \mathbb{R}^n$  such that*

$$\langle s, x \rangle > \sup_{y \in C} \langle s, y \rangle. \quad (4.1.1)$$

PROOF. Set  $s := x - p_C(x) \neq 0$ . We write (3.1.3) as

$$0 \geq \langle s, y - x + s \rangle = \langle s, y \rangle - \langle s, x \rangle + \|s\|^2.$$

Thus we have

$$\langle s, x \rangle - \|s\|^2 \geq \langle s, y \rangle \quad \text{for all } y \in C,$$

and our  $s$  is a convenient answer for (4.1.1).  $\square$

Naturally,  $s$  could be replaced by  $-s$  in (4.1.1) and Theorem 4.1.1 could just be stated as: there exists  $s' \in \mathbb{R}^n$  such that

$$\langle s', x \rangle < \inf \{\langle s', y \rangle : y \in C\}.$$

Geometrically, we know that an  $s \neq 0$  defines hyperplanes  $H_{s,r}$  as in Example 1.1.2(a), which are translations of each other when  $r$  describes  $\mathbb{R}$ . With  $s$  of (4.1.1) (which is certainly nonzero!), pick

$$r = r_s := \frac{1}{2}(\langle s, x \rangle + \sup_{y \in C} \langle y, s \rangle).$$

Then

$$\langle s, x \rangle - r_s > 0 \quad \text{and} \quad \langle s, y \rangle - r_s < 0 \quad \text{for all } y \in C,$$

which can be summarized in one sentence: the affine hyperplane  $H_{s,r_s}$  separates the two convex sets  $C$  and  $\{x\}$ . These two sets are in the opposite (open) half-spaces limited by that hyperplane.

**Remark 4.1.2** With relation to this interpretation, Theorem 4.1.1 is often called the *Hahn-Banach Theorem* in geometric form. On the other hand, consider the right-hand side of (4.1.1); it suggests a function  $\sigma_C : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ , called the *support function* of  $C$ :

$$\sigma_C(s) := \sup \{\langle s, y \rangle : y \in C\},$$

which will be studied thoroughly in Chap. V. If  $x \in C$ , we have by definition

$$\langle s, x \rangle \leq \sigma_C(s) \quad \text{for all } s \in \mathbb{R}^n;$$

but this actually *characterizes* the elements of  $C$ : Theorem 4.1.1 tells us that the converse is true. Therefore the test “ $x \in C$ ?” is equivalent to the test “ $\langle \cdot, x \rangle \leq \sigma_C$ ?”, which compares the linear function  $\langle \cdot, x \rangle$  to the function  $\sigma_C$ . With this interpretation, Theorem 4.1.1 can be formulated in an equivalent analytical way, involving functions instead of hyperplanes; this is called the Hahn-Banach Theorem in *analytical* form.  $\square$

A convenient generalization of Theorem 4.1.1 is the following:

**Corollary 4.1.3 (Strict Separation of Convex Sets)** *Let  $C_1, C_2$  be two nonempty closed convex sets with  $C_1 \cap C_2 = \emptyset$ . If  $C_2$  is bounded, there exists  $s \in \mathbb{R}^n$  such that*

$$\sup_{y \in C_1} \langle s, y \rangle < \min_{y \in C_2} \langle s, y \rangle. \tag{4.1.2}$$

PROOF. The set  $C_1 - C_2$  is convex (Proposition 1.2.4) and closed (because  $C_2$  is compact). To say that  $C_1$  and  $C_2$  are disjoint is to say that  $0 \notin C_1 - C_2$ , so we have by Theorem 4.1.1 an  $s \in \mathbb{R}^n$  separating  $\{0\}$  from  $C_1 - C_2$ :

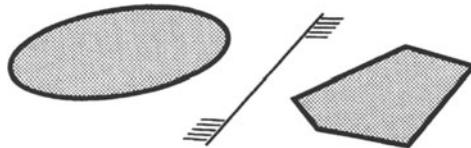
$$\sup \{\langle s, y \rangle : y \in C_1 - C_2\} < \langle s, 0 \rangle = 0.$$

This means:

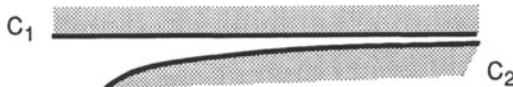
$$\begin{aligned} 0 &> \sup_{y_1 \in C_1} \langle s, y_1 \rangle + \sup_{y_2 \in C_2} \langle s, -y_2 \rangle \\ &= \sup_{y_1 \in C_1} \langle s, y_1 \rangle - \inf_{y_2 \in C_2} \langle s, y_2 \rangle. \end{aligned}$$

Because  $C_2$  is bounded, the last infimum (is a min and) is finite and can be moved to the left-hand side.  $\square$

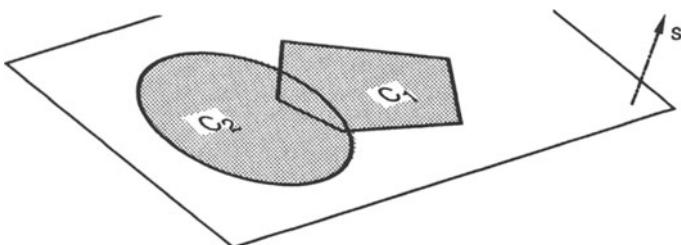
Once again, (4.1.2) can be switched over to  $\inf_{y \in C_1} \langle s, y \rangle > \max_{y \in C_2} \langle s, y \rangle$ . Using the support function of Remark 4.1.2, we can also write (4.1.2) as  $\sigma_{C_1}(s) + \sigma_{C_2}(-s) < 0$ . Figure 4.1.1 gives the same geometric interpretation as before. Choosing  $r = r_s$  strictly between  $\sigma_{C_1}(s)$  and  $-\sigma_{C_2}(-s)$ , we obtain a hyperplane separating  $C_1$  and  $C_2$  strictly: each set is in one of the corresponding open half-spaces.



**Fig. 4.1.1.** Strict separation of two convex sets



**Fig. 4.1.2.** Strict separation needs compactness



**Fig. 4.1.3.** An improper separation

When  $C_1$  and  $C_2$  are both unbounded, Corollary 4.1.3 may fail – even though the role of boundedness was apparently minor, but see Fig. 4.1.2. As suggested by this picture,  $C_1$  and

$C_2$  can nevertheless be *weakly* separated, i.e. (4.1.2) can be replaced by a weak inequality. Such a weakening is a bit exaggerated, however: Fig. 4.1.3 shows that (4.1.2) may hold as

$$\langle s, y_1 \rangle = \langle s, y_2 \rangle \quad \text{for all } (y_1, y_2) \in C_1 \times C_2$$

if  $s$  is orthogonal to  $\text{aff}(C_1 \cup C_2)$ . For a convenient definition, we need to be more demanding: we say that the two nonempty convex sets  $C_1$  and  $C_2$  are *properly separated* by  $s \in \mathbb{R}^n$  when

$$\sup_{y_1 \in C_1} \langle s, y_1 \rangle \leq \inf_{y_2 \in C_2} \langle s, y_2 \rangle \quad \text{and} \quad \inf_{y_1 \in C_1} \langle s, y_1 \rangle < \sup_{y_2 \in C_2} \langle s, y_2 \rangle.$$

This (weak) proper separation property is sometimes just what is needed for technical purposes. It happens to hold under fairly general assumptions on the intersection  $C_1 \cap C_2$ . We end this section with a possible result, stated without proof.

**Theorem 4.1.4 (Proper Separation of Convex Sets)** *If the two nonempty convex sets  $C_1$  and  $C_2$  satisfy  $(\text{ri } C_1) \cap (\text{ri } C_2) = \emptyset$ , they can be properly separated.*  $\square$

Observe the assumption coming into play. We have already seen it in Proposition 2.1.10, and we know from (2.1.5) that it is equivalent to

$$0 \notin \text{ri}(C_1 - C_2).$$

## 4.2 First Consequences of the Separation Properties

The separation properties introduced in §4.1 have many applications. To prove that some set  $S$  is contained in a closed convex set  $C$ , a possibility is often to argue by contradiction, separating from  $C$  a point in  $S \setminus C$ , and then exploiting the simple structure of the separating hyperplane. Here we review some of these applications, including the proofs announced in the previous sections. Note: our proofs are often fairly short (as is that of Corollary 4.1.3) or geometrical. It is a good exercise to develop more elementary proofs, or to support the geometry with detailed calculations.

**(a) Existence of Supporting Hyperplanes** First of all, we note that a convex set  $C$ , not equal to the whole of  $\mathbb{R}^n$ , does have a supporting hyperplane in the sense of Definition 2.4.1. To see it, use first Proposition 2.1.8:  $\text{cl } C \neq \mathbb{R}^n$  (otherwise, we would have the contradiction  $C \supset \text{ri } C = \text{ri cl } C = \text{ri } \mathbb{R}^n = \mathbb{R}^n$ ). Then take a hyperplane separating  $\text{cl } C$  from some  $x \notin \text{cl } C$ : it is our asserted support of  $C$ . Actually, we can prove slightly more:

**Lemma 4.2.1** *Let  $x \in \text{bd } C$ , where  $C \neq \emptyset$  is convex in  $\mathbb{R}^n$  (naturally  $C \neq \mathbb{R}^n$ ). There exists a hyperplane supporting  $C$  at  $x$ .*

PROOF. Because  $C$ ,  $\text{cl } C$  and their complements have the same boundary (remember Remark 2.1.9), a sequence  $\{x_k\}$  can be found such that

$$x_k \notin \text{cl } C \quad \text{for } k = 1, 2, \dots \quad \text{and} \quad \lim_{k \rightarrow +\infty} x_k = x.$$

For each  $k$  we have by Theorem 4.1.1 some  $s_k$  with  $\|s_k\| = 1$  such that

$$\langle s_k, x_k - y \rangle > 0 \quad \text{for all } y \in C \subset \text{cl } C.$$

Extract a subsequence if necessary so that  $s_k \rightarrow s$  (note:  $s \neq 0$ ) and pass to the limit to obtain

$$\langle s, x - y \rangle \geq 0 \quad \text{for all } y \in C,$$

which is the required result  $\langle s, x \rangle = r \geq \langle s, y \rangle$  for all  $y \in C$ .  $\square$

**Remark 4.2.2** The above procedure may well end up with a supporting hyperplane containing  $C$ :  $\langle s, x - y \rangle = 0$  for all  $y \in C$ , a result of little interest; see also Fig. 4.1.3. This can happen only when  $C$  is a “flat” convex set ( $\dim C \leq n - 1$ ), in which case our construction should be done in  $\text{aff } C$ , as illustrated on Fig. 4.2.1. Let us detail such a “relative” construction, to demonstrate a calculation involving affine hulls.

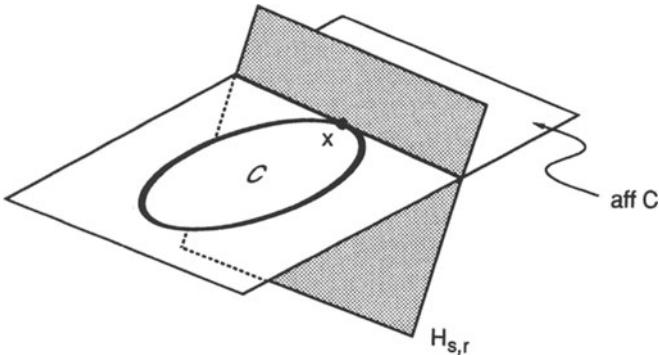


Fig. 4.2.1. Nontrivial supports

Let  $V$  be the subspace parallel to  $\text{aff } C$ , with  $U = V^\perp$  its orthogonal subspace: by definition,  $\langle s, y - x \rangle = 0$  for all  $s \in U$  and  $y \in C$ . Suppose  $x \in \text{rbd } C$  (the case  $x \in \text{ri } C$  is hopeless) and translate  $C$  to  $C_0 := C - \{x\}$ . Then  $C_0$  is a convex set in the Euclidean space  $V$  and  $0 \in \text{rbd } C_0$ . We take as in 4.2.1 a sequence  $\{x_k\} \subset V \setminus \text{cl } C_0$  tending to 0 and a corresponding unitary  $s_k \in V$  separating the point  $x_k$  from  $C_0$ . The limit  $s \neq 0$  is in  $V$ , separates (not strictly)  $\{0\}$  and  $C_0$ , i.e.  $\{x\}$  and  $C$ : we are done.

We will say that  $H_{s,r}$  is a *nontrivial support* (at  $x$ ) if  $s \notin U$ , i.e. if  $s_V \neq 0$ , with the decomposition  $s = s_V + s_U$ . Then  $C$  is not contained in  $H_{s,r}$ : if it were, we would have for all  $y \in C$

$$r = \langle s, y \rangle = \langle s_V, y \rangle + \langle s_U, x \rangle.$$

In other words,  $\langle s_V, \cdot \rangle$  would be constant on  $C$ ; by definition of the affine hull and of  $V$ , this would mean  $s_V \in U$ , i.e. the contradiction  $s_V = 0$ . To finish, note that  $s_U$  may be assumed to be 0: if  $s_V + s_U$  is a nontrivial support, so is  $s_V = s_V + 0$  as well; it corresponds to a hyperplane orthogonal to  $C$ .  $\square$

In terms of Carathéodory’s Theorem 1.3.6, a consequence of our existence lemma is the following:

**Proposition 4.2.3** *Let  $S \subset \mathbb{R}^n$  and  $C := \text{co } S$ . Any  $x \in C \cap \text{bd } C$  can be represented as a convex combination of  $n$  elements of  $S$ .*

PROOF. Because  $x \in \text{bd } C$ , there is a hyperplane  $H_{s,r}$  supporting  $C$  at  $x$ : for some  $s \neq 0$  and  $r \in \mathbb{R}$ ,

$$\langle s, x \rangle - r = 0 \quad \text{and} \quad \langle s, y \rangle - r \leq 0 \quad \text{for all } y \in C. \quad (4.2.1)$$

On the other hand, Carathéodory's Theorem 1.3.6 implies the existence of points  $x_1, \dots, x_{n+1}$  in  $S$  and convex multipliers  $\alpha_1, \dots, \alpha_{n+1}$  such that  $x = \sum_{i=1}^{n+1} \alpha_i x_i$ ; and each  $\alpha_i$  can be assumed positive (otherwise the proof is finished).

Setting successively  $y = x_i$  in (4.2.1), we obtain by convex combination

$$0 = \langle s, x \rangle - r = \sum_{i=1}^{n+1} \alpha_i (\langle s, x_i \rangle - r) \leq 0,$$

so each  $\langle s, x_i \rangle - r$  is actually 0. Each  $x_i$  is therefore not only in  $S$ , but also in  $H_{s,r}$ , a set whose dimension is  $n - 1$ . It follows that our starting  $x$ , which is in  $\text{co}\{x_1, \dots, x_{n+1}\}$ , can be described as the convex hull of only  $n$  among these  $x_i$ 's.  $\square$

**(b) Outer Description of Closed Convex Sets** Closing a (convex) set consists in intersecting the closed (convex) sets containing it. We mentioned in Remark 1.4.4 that convexity allowed the intersection to be restricted to a simple class of closed convex sets: the closed half-spaces. Indeed, Lemma 4.2.1 ensures that a nonempty convex set  $C \subsetneq \mathbb{R}^n$  has at least one supporting hyperplane: if we denote by

$$H_{s,r}^- := \{y \in \mathbb{R}^n : \langle s, y \rangle \leq r\}$$

a closed half-space defined by a given  $(s, r) \in \mathbb{R}^n \times \mathbb{R}$ , ( $s \neq 0$ ), then the index-set

$$\begin{aligned} \Sigma_C &:= \{(s, r) \in \mathbb{R}^n \times \mathbb{R} : C \subset H_{s,r}^-\} \\ &= \{(s, r) : \langle s, y \rangle \leq r \text{ for all } y \in C\} \end{aligned} \quad (4.2.2)$$

is nonempty. As illustrated by Fig. 4.2.2, we can therefore intersect all the half-spaces indexed in  $\Sigma_C$ :

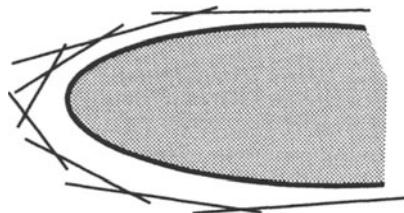


Fig. 4.2.2. Outer construction of a closed convex set

$$\begin{aligned} C \subset C^* &:= \cap_{(s,r) \in \Sigma_C} H_{s,r}^- = \\ &\{z \in \mathbb{R}^n : \langle s, z \rangle \leq r \text{ whenever } \langle s, y \rangle \leq r \text{ for all } y \in C\}. \end{aligned}$$

**Theorem 4.2.4** Let  $\emptyset \neq C \subsetneq \mathbb{R}^n$  be convex. The set  $C^*$  defined above is the closure of  $C$ .

PROOF. By construction,  $C^* \supset \text{cl } C$ . Conversely, take  $x \notin \text{cl } C$ ; we can separate  $x$  and  $\text{cl } C$ : there exists  $s_0 \neq 0$  such that

$$\langle s_0, x \rangle > \sup_{y \in C} \langle s_0, y \rangle =: r_0.$$

Then  $(s_0, r_0) \in \Sigma_C$ ; but  $x \notin H_{s_0, r_0}^-$ , hence  $x \notin C^*$ .  $\square$

The definition of  $C^*$ , rather involved, can be slightly simplified: actually,  $\Sigma_C$  is redundant, as it contains much too many  $r$ 's. Roughly speaking, for given  $s \in \mathbb{R}^n$ , just take the number

$$r = r_s := \inf \{r \in \mathbb{R} : (s, r) \in \Sigma_C\}$$

that is sharp in (4.2.2). Letting  $s$  vary,  $(s, r_s)$  describes a set  $\Sigma_C^0$ , smaller than  $\Sigma_C$  but just as useful. With this new notation, the expression of  $C^* = \text{cl } C$  reduces to

$$\text{cl } C = \{z \in \mathbb{R}^n : \langle s, z \rangle \leq \sup_{y \in C} \langle s, y \rangle\}.$$

We find again the support function of Remark 4.1.2 coming into play. Chapter V will follow this development more thoroughly.

The message from Theorem 4.2.4 is that a closed convex set can thus be *defined* as the intersection of the closed half-spaces containing it:

**Corollary 4.2.5** The data  $(s_j, r_j) \in \mathbb{R}^n \times \mathbb{R}$  for  $j$  in an arbitrary index set  $J$  is equivalent to the data of a closed convex set  $C$  via the relation

$$C = \bigcap_{j \in J} \{x \in \mathbb{R}^n : \langle s_j, x \rangle \leq r_j\}.$$

PROOF. If  $C$  is given, define  $\{(s_j, r_j)\}_J := \Sigma_C$  as in (4.2.2). If  $\{(s_j, r_j)\}_J$  is given, the intersection of the corresponding half-spaces is a closed convex set. Note here that we can define at the same time the whole of  $\mathbb{R}^n$  and the empty sets as two extreme cases.  $\square$

As an important special case, we find:

**Definition 4.2.6 (Polyhedral Sets)** A closed convex polyhedron is an intersection of finitely many half-spaces. Take  $(s_1, r_1), \dots, (s_m, r_m)$  in  $\mathbb{R}^n \times \mathbb{R}$ , with  $s_i \neq 0$  for  $i = 1, \dots, m$ ; then define

$$P := \{x \in \mathbb{R}^n : \langle s_j, x \rangle \leq r_j \text{ for } j = 1, \dots, m\},$$

or in matrix notations (assuming the dot-product for  $\langle \cdot, \cdot \rangle$ ),

$$P = \{x \in \mathbb{R}^n : Ax \leq b\},$$

if  $A$  is the matrix whose rows are  $s_j$  and  $b \in \mathbb{R}^m$  has coordinates  $r_j$ .

A closed convex polyhedral cone is the special case where  $b = 0$ .  $\square$

**(c) Proof of Minkowski's Theorem** We turn now to the inner description of a convex set and prove Theorem 2.3.4, asserting that  $C = \text{co ext } C$  when  $C$  is compact convex.

The result is trivially true if  $\dim C = 0$ , i.e.  $C$  is a singleton, with a unique extreme point. Assume for induction that the result is true for compact convex sets of dimension less than  $k$ ; let  $C$  be a compact convex set of dimension  $k$  and take  $x \in C$ . There are two possibilities:

- If  $x \in \text{rbd } C$ , §4.2(a) tells us that there exists a nontrivial hyperplane  $H$  supporting  $C$  at  $x$ . The nonempty compact convex set  $C \cap H$  has dimension at most  $k - 1$ , so  $x \in C \cap H$  is a convex combination of extreme points in that set, which is an exposed face of  $C$ . Using Remark 2.4.4, these extreme points are also extreme in  $C$ .
- If  $x \in \text{ri } C (= C \setminus \text{rbd } C)$ , take in  $C$  a point  $x' \neq x$ ; this is possible for  $\dim C > 0$ . The affine line generated by  $x$  and  $x'$  cuts  $\text{rbd } C$  in at most two points  $y$  and  $z$  (see Remark 2.1.7, there are really two points because  $C$  is compact). From the first part of the proof,  $y$  and  $z$  are convex combinations of extreme points in  $C$ ; and so is their convex combination  $x$  (associativity of convex combinations).

**(d) Bipolar of a Convex Cone** The definition of a polar cone was given in §3.2, where some interesting properties were pointed out. Here we can show one more similarity with the concept of orthogonality in linear analysis.

**Proposition 4.2.7** *Let  $K$  be a convex cone with polar  $K^\circ$ ; then, the polar  $K^{\circ\circ}$  of  $K^\circ$  is the closure of  $K$ .*

PROOF. We exploit Remark 4.1.2: due to its conical character ( $\alpha x \in K$  if  $x \in K$  and  $\alpha > 0$ ),  $\text{cl } K$  has a very special support function:

$$\sigma_{\text{cl } K}(s) = \begin{cases} \langle s, 0 \rangle = 0 & \text{if } \langle s, x \rangle \leq 0 \text{ for all } x \in \text{cl } K, \\ +\infty & \text{otherwise.} \end{cases}$$

In other words,  $\sigma_{\text{cl } K}$  is 0 on  $K^\circ$ ,  $+\infty$  elsewhere. Thus, the characterization

$$x \in \text{cl } K \iff \langle \cdot, x \rangle \leq \sigma_{\text{cl } K}(\cdot)$$

becomes

$$x \in \text{cl } K \iff \begin{cases} \langle s, x \rangle \leq 0 & \text{for all } s \in K^\circ \\ (\langle s, x \rangle \text{ arbitrary} & \text{for } s \notin K^\circ!) \end{cases},$$

in which we recognize the definition of  $K^{\circ\circ}$ . □

Of course, if  $K$  is already closed,  $K^{\circ\circ} = K$ . With relation to (a) above, we observe that every supporting hyperplane of  $K$  at  $x \in \text{bd } K$  also supports  $K$  at 0: when dealing with supports to a cone, it is enough to consider *linear* hyperplanes only.

**Remark 4.2.8** Consider the index-set  $\Sigma_K$  of (4.2.2), associated with a closed convex cone  $K$ : its  $r$ -part can be restricted to  $\{0\}$ ; as for its  $s$ -part, we see from Definition 3.2.1 that it becomes  $K^\circ \setminus \{0\}$ . In other words: barring the zero-vector, a closed convex cone is the set of (linear) hyperplanes supporting its polar at 0. □

### 4.3 The Lemma of Minkowski-Farkas

Because of its historical importance, we devote an entire subsection to another consequence of the separation property, known as Farkas' Lemma. Let us first recall a classical result from linear algebra: if  $A$  is a matrix with  $n$  rows and  $m$  columns and  $b \in \mathbb{R}^n$ , the system  $A\alpha = b$  has a solution in  $\mathbb{R}^m$  (we say that the system is *consistent*) exactly when

$$b \in \text{Im } A = [\text{Ker } A^\top]^\perp ;$$

this can be rewritten  $\{b\}^\perp \supset \text{Ker } A^\top$ , or

$$\{x \in \mathbb{R}^n : A^\top x = 0\} \subset \{x \in \mathbb{R}^n : b^\top x = 0\} .$$

Denoting by  $s_1, \dots, s_m$  the columns of  $A$  and using our Euclidean notation, we write the equivalence of these properties as

$$\begin{aligned} b \in \text{lin}\{s_1, \dots, s_m\} &\quad \text{if and only if} \\ \langle b, x \rangle = 0 &\quad \text{whenever } \langle s_j, x \rangle = 0 \text{ for } j = 1, \dots, m . \end{aligned}$$

Moving to the unilateral world of convex analysis, we replace linear hulls by conical hulls, and equalities by inequalities. This gives a result dating back to the end of the XIX<sup>th</sup> Century, due to J. Farkas and also to H. Minkowski; we state it without proof, as it will be a consequence of Theorem 4.3.4 below.

**Lemma 4.3.1 (Farkas I)** *Let  $b, s_1, \dots, s_m$  be given in  $\mathbb{R}^n$ . The set*

$$\{x \in \mathbb{R}^n : \langle s_j, x \rangle \leq 0 \text{ for } j = 1, \dots, m\} \tag{4.3.1}$$

*is contained in the set*

$$\{x \in \mathbb{R}^n : \langle b, x \rangle \leq 0\} \tag{4.3.2}$$

*if and only if (see Definition 1.4.5 of a conical hull)*

$$b \in \text{cone}\{s_1, \dots, s_m\} . \tag{4.3.3}$$

□

To express the inclusion relation between the sets (4.3.1) and (4.3.2), one also says that the inequality with  $b$  is a *consequence* of the joint inequalities with  $s_j$ . Another way of expressing (4.3.3) is to say that the system of equations and inequations in  $\alpha$

$$b = \sum_{j=1}^m \alpha_j s_j, \quad \alpha_j \geq 0 \text{ for } j = 1, \dots, m \tag{4.3.4}$$

has a solution.

Farkas' Lemma is sometimes formulated as an *alternative*, i.e. a set of two statements such that each one is false when the other is true. More precisely, let  $P$  and  $Q$  be two logical propositions. They are said to form an alternative if one and only one of them is true:

$$P \implies \text{not } Q \quad \text{and} \quad \text{not } P \implies Q$$

or, just as simply:

$$P \iff \text{not } Q \quad [\text{or} \quad Q \iff \text{not } P].$$

This applies to Farkas' lemma:

**Lemma 4.3.2 (Farkas II)** *Let  $b, s_1, \dots, s_m$  be given in  $\mathbb{R}^n$ . Then exactly one of the following statements is true.*

$$P := (4.3.4) \text{ has a solution } \alpha \in \mathbb{R}^n.$$

$$Q := \begin{cases} \text{The system of inequations} \\ \langle b, x \rangle > 0, \quad \langle s_j, x \rangle \leq 0 \text{ for } j = 1, \dots, m \\ \text{has a solution } x \in \mathbb{R}^n. \end{cases}$$

□

Still another formulation is geometric. Call  $K$  the convex cone generated by  $s_1, \dots, s_m$ ; as seen in Example 3.2.2,  $K^\circ$  is the set (4.3.1). What Farkas' Lemma says is that

$$\begin{aligned} b \in K & \text{ [i.e. (4.3.3) holds] if and only if} \\ & \langle b, x \rangle \leq 0 \text{ whenever } x \in K^\circ \text{ [i.e. } b \in K^{\circ\circ}\text{].} \end{aligned}$$

More simply, Farkas' Lemma is:  $K^{\circ\circ} = K$ ; but we know from §4.2(d) that this property holds under the sole condition that  $K$  is closed. The proof of Farkas' Lemma therefore reduces to proving the following result:

**Lemma 4.3.3 (Farkas III)** *Let  $s_1, \dots, s_m$  be given in  $\mathbb{R}^n$ . Then the convex cone*

$$K := \text{cone}\{s_1, \dots, s_m\} = \left\{ \sum_{j=1}^m \alpha_j s_j : \alpha_j \geq 0 \text{ for } j = 1, \dots, m \right\}$$

*is closed.*

PROOF. It is quite similar to that of Carathéodory's Theorem 1.3.6. First, the proof is easy if the  $s_j$ 's are linearly independent: then, the convergence of

$$x^k = \sum_{j=1}^m \alpha_j^k s_j \quad \text{for } k \rightarrow \infty \tag{4.3.5}$$

is equivalent to the convergence of each  $\{\alpha_j^k\}$  to some  $\alpha_j$ , which must be nonnegative if each  $\alpha_j^k$  in (4.3.5) is nonnegative.

Suppose, on the contrary, that the system  $\sum_{j=1}^m \beta_j s_j = 0$  has a nonzero solution  $\beta \in \mathbb{R}^m$  and assume  $\beta_j < 0$  for some  $j$  (change  $\beta$  to  $-\beta$  if necessary). As in the proof of Theorem 1.3.6, write each  $x \in K$  as

$$x = \sum_{j=1}^m \alpha_j s_j = \sum_{j=1}^m [\alpha_j + t^*(x)\beta_j] s_j = \sum_{j \neq i(x)} \alpha'_j s_j,$$

where

$$i(x) \in \operatorname{Argmin}_{\beta_j < 0} \frac{-\alpha_j}{\beta_j}, \quad t^*(x) := \frac{-\alpha_{i(x)}}{\beta_{i(x)}},$$

so that each  $\alpha'_j = \alpha_j + t^*(x)\beta_j$  is nonnegative. Letting  $x$  vary in  $K$ , we thus construct a decomposition

$$K = \cup\{K_i : i = 1, \dots, m\},$$

where  $K_i$  is the conical hull of the  $m - 1$  generators  $s_j$ ,  $j \neq i$ .

Now, if there is some  $i$  such that the generators of  $K_i$  are linearly dependent, we repeat the argument for a further decomposition of this  $K_i$ . After finitely many such operations, we end up with a decomposition of  $K$  as a finite union of polyhedral convex cones, each having linearly independent generators. All these cones are therefore closed (first part of the proof), so  $K$  is closed as well.  $\square$

We are now in a position to state a general version of Farkas' Lemma, with non-homogeneous terms and infinitely many inequalities. Its proof uses in a direct way the separation Theorem 4.1.1.

**Theorem 4.3.4 (Generalized Farkas)** *Let be given  $(b, r)$  and  $(s_j, \rho_j)$  in  $\mathbb{R}^n \times \mathbb{R}$ , where  $j$  varies in an (arbitrary) index set  $J$ . Suppose that the system of inequalities*

$$\langle s_j, x \rangle \leq \rho_j \quad \text{for all } j \in J \tag{4.3.6}$$

*has a solution  $x \in \mathbb{R}^n$  (the system is consistent). Then the following two properties are equivalent:*

- (i)  $\langle b, x \rangle \leq r$  for all  $x$  satisfying (4.3.6);
- (ii)  $(b, r)$  is in the closed convex conical hull of  $S := \{(0, 1)\} \cup \{(s_j, \rho_j)\}_{j \in J}$ .

PROOF. [(ii)  $\Rightarrow$  (i)] Let first  $(b, r)$  be in  $K := \text{cone } S$ . In other words, there exists a finite set  $\{1, \dots, m\} \subset J$  and nonnegative  $\alpha_0, \alpha_1, \dots, \alpha_m$  such that (we adopt the convention  $\sum_{\emptyset} = 0$ )

$$b = \sum_{j=1}^m \alpha_j s_j \quad \text{and} \quad r = \alpha_0 + \sum_{j=1}^m \alpha_j \rho_j.$$

For each  $x$  satisfying (4.3.6) we can write

$$\langle b, x \rangle \leq r - \alpha_0 \leq r. \tag{4.3.7}$$

If, now,  $(b, r)$  is in the closure of  $K$ , pass to the limit in (4.3.7) to establish the required conclusion (i) for all  $(b, r)$  described by (ii).

[(i)  $\Rightarrow$  (ii)] If  $(b, r) \notin \text{cl } K$ , separate  $(b, r)$  from  $\text{cl } K$ : equipping  $\mathbb{R}^n \times \mathbb{R}$  with the scalar product

$$\langle\langle (b, r), (d, t) \rangle\rangle := \langle b, d \rangle + rt,$$

there exists  $(d, -t) \in \mathbb{R}^n \times \mathbb{R}$  such that

$$\sup_{(s,\rho) \in K} [\langle s, d \rangle - \rho t] < \langle b, d \rangle - rt. \quad (4.3.8)$$

It follows first that the left-hand supremum is a finite number  $\kappa$ . Then the conical character of  $K$  implies  $\kappa \leq 0$ , because  $\alpha\kappa \leq \kappa$  for all  $\alpha > 0$ ; actually  $\kappa = 0$  because  $(0, 0) \in K$ . In summary, we have singled out  $(d, t) \in \mathbb{R}^n \times \mathbb{R}$  such that

$$\begin{aligned} t &\geq 0 && [\text{take } (0, 1) \in K] \\ (*) \quad \langle s_j, d \rangle - \rho_j t &\leq 0 \text{ for all } j \in J && [\text{take } (s_j, \rho_j) \in K] \\ (**)\quad \langle b, d \rangle - rt &> 0. && [\text{don't forget (4.3.8)}] \end{aligned}$$

Now consider two cases:

- If  $t > 0$ , divide  $(*)$  and  $(**)$  by  $t$  to exhibit the point  $x = d/t$  violating (i).
- If  $t = 0$ , take  $x_0$  satisfying (4.3.6). Observe from  $(*)$  that, for all  $\alpha > 0$ , the point  $x(\alpha) = x_0 + \alpha d$  satisfies (4.3.6) as well. Yet, let  $\alpha \rightarrow +\infty$  in

$$\langle b, x(\alpha) \rangle = \langle b, x_0 \rangle + \alpha \langle b, d \rangle$$

to realize from  $(**)$  that  $x(\alpha)$  violates (i) if  $\alpha$  is large enough.

Thus we have proved in both cases that “not (ii)  $\Rightarrow$  not (i)”.  $\square$

We finish with two comments relating Theorem 4.3.4 with the previous forms of Farkas’ Lemma. Take first the homogeneous case, where  $r$  and the  $\rho_j$ ’s are all zero. Then the consistency assumption is automatically satisfied (by  $x = 0$ ) and the theorem says:

$$(i') [\langle s_j, x \rangle \leq 0 \text{ for } j \in J] \implies [\langle b, x \rangle \leq 0]$$

is equivalent to

$$(ii') b \in \overline{\text{cone}}\{s_j : j \in J\}.$$

Second, suppose that  $J = \{1, \dots, m\}$  is a finite set, so the set described by (4.3.6) becomes a closed convex polyhedron, assumed nonempty. A handy matrix notation (assuming the dot-product for  $\langle \cdot, \cdot \rangle$ ), is  $A^\top x \leq \rho$ , if  $A$  is the matrix whose columns are the  $s_j$ ’s, and  $\rho \in \mathbb{R}^m$  has the coordinates  $\rho_1, \dots, \rho_m$ . Then Theorem 4.3.4 writes:

$$(i'') \{x \in \mathbb{R}^n : A^\top x \leq \rho\} \subset \{x \in \mathbb{R}^n : b^\top x \leq r\}$$

is equivalent to

$$(ii'') \exists \alpha \in \mathbb{R}^m \text{ such that } \alpha \geq 0, A\alpha = b, \rho^\top \alpha \leq r.$$

Indeed, it suffices to recall Lemma 4.3.3: the conical hull involved in (ii) of Theorem 4.3.4 is already closed. Beware that the last relation in (ii'') is really an inequality.

## 5 Conical Approximations of Convex Sets

Given a set  $S$  and  $x \in S$ , a fruitful idea is to approximate  $S$  near  $x$  by a “simpler” set. In classical differential geometry, a “smooth” surface  $S$  is approximated by an

affine manifold “tangent” to  $S$ . This concept is most exploited in the differentiation of a “smooth” function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , whose graph is “tangent” to an affine hyperplane in  $\mathbb{R}^n \times \mathbb{R}$ :

$$\text{gr } f \simeq \{(y, r) : r - f(x) = \langle \nabla f(x), y - x \rangle\}.$$

Because convex sets have no reason to be “smooth”, some substitute to affine manifolds must be proposed. We know that affine manifolds are translations of subspaces; say, we approximate  $S$  near  $x$  by

$$S \simeq H_S(x) = \{x\} + V_S(x),$$

where  $V_S(x)$  is a subspace: the subspace tangent to  $S$  at  $x$ . It is therefore time to remember §3.2: in convex analysis, the natural substitutes for subspaces are the *closed convex cones*. Besides, another important object is the set of normals to  $S$  at  $x$ , i.e. the subspace orthogonal to  $V_S(x)$ ; here, orthogonality will be replaced by *polarity*, as in Moreau’s Theorem 3.2.5.

## 5.1 Convenient Definitions of Tangent Cones

In order to introduce the convenient objects, we need first to cast a fresh glance at the general concept of tangency. We therefore consider in this subsection an arbitrary *closed* subset  $S \subset \mathbb{R}^n$ .

A direction  $d$  is classically called tangent to  $S$  at  $x \in S$  when it is the derivative at  $x$  of some curve drawn on  $S$ ; it follows that  $-d$  is a tangent as well. Since we are rather interested by cones, we will simply require a *half-derivative* from the curve in question – incidentally, taking half-derivatives goes together with §I.4.1. Furthermore, sets of discrete type cannot have any tangent direction in the above sense, we will therefore replace *curves* by *sequences*. In a word, our new definition of tangency is as follows:

**Definition 5.1.1** Let  $S \subset \mathbb{R}^n$  be nonempty. We say that  $d \in \mathbb{R}^n$  is a direction *tangent* to  $S$  at  $x \in S$  when there exists a sequence  $\{x_k\} \subset S$  and a sequence  $\{t_k\}$  such that, when  $k \rightarrow +\infty$ ,

$$x_k \rightarrow x, \quad t_k \downarrow 0, \quad \frac{x_k - x}{t_k} \rightarrow d. \quad (5.1.1)$$

The set of all such directions is called the *tangent cone* (also called the contingent cone, or Bouligand’s cone) to  $S$  at  $x \in S$ , denoted by  $T_S(x)$ .  $\square$

Observe immediately that  $0$  is always a tangent direction (take  $x_k \equiv x$ !); also, if  $d$  is tangent, so is  $\alpha d$  for any  $\alpha > 0$  (change  $t_k$  to  $t_k/\alpha$ !). The terminology “tangent cone” is therefore legal. If  $x \in \text{int } S$ ,  $T_S(x)$  is clearly the whole space, so that the only interesting points are those on  $\text{bd } S$ .

If we set in Definition 5.1.1  $d_k := (x_k - x)/t_k \rightarrow d$ , i.e.  $x_k = x + t_k d_k$  [ $\in S$ ], we obtain the equivalent formulation:

**Proposition 5.1.2** *A direction  $d$  is tangent to  $S$  at  $x \in S$  if and only if*

$$\exists \{d_k\} \rightarrow d, \exists \{t_k\} \downarrow 0 \text{ such that } x + t_k d_k \in S \text{ for all } k.$$

□

A tangent direction thus appears as a set of limits; a limit of tangent directions is therefore a “limit of limits”, and is a limit itself:

**Proposition 5.1.3** *The tangent cone is closed.*

PROOF. Let  $\{d_\ell\} \subset T_S(x)$  be converging to  $d$ ; for each  $\ell$  take sequences  $\{x_{\ell,k}\}_k$  and  $\{t_{\ell,k}\}_k$  associated with  $d_\ell$  in the sense of Definition 5.1.1. Fix  $\ell > 0$ : we can find  $k_\ell$  such that

$$\left\| \frac{x_{\ell,k_\ell} - x}{t_{\ell,k_\ell}} - d_\ell \right\| \leq \frac{1}{\ell}.$$

Letting  $\ell \rightarrow \infty$ , we then obtain the sequences  $\{x_{\ell,k_\ell}\}_\ell$  and  $\{t_{\ell,k_\ell}\}_\ell$  which define  $d$  as an element of  $T_S(x)$ . □

The examples below confirm that our definition reproduces the classical one when  $S$  is “well-behaved”, while Fig. 5.1.1 illustrates a case where classical tangency cannot be used.

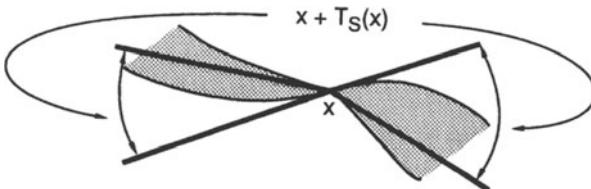


Fig. 5.1.1. Tangency to a “bad” set

**Examples 5.1.4** Given  $m$  functions  $c_1, \dots, c_m$  continuously differentiable on  $\mathbb{R}^n$ , consider

$$S := \{x \in \mathbb{R}^n : c_i(x) = 0 \text{ for } i = 1, \dots, m\}.$$

Let  $x \in S$  be such that the gradients  $\nabla c_1(x), \dots, \nabla c_m(x)$  are linearly independent. Then  $T_S(x)$  is the subspace

$$\{d \in \mathbb{R}^n : \langle \nabla c_i(x), d \rangle = 0 \text{ for } i = 1, \dots, m\}. \quad (5.1.2)$$

Another example is

$$S := \{x \in \mathbb{R}^n : c_1(x) \leq 0\}.$$

At  $x \in S$  such that  $c_1(x) = 0$  and  $\nabla c_1(x) \neq 0$ ,  $T_S(x)$  is the *half-space*

$$\{d \in \mathbb{R}^n : \langle \nabla c_1(x), d \rangle \leq 0\}. \quad (5.1.3)$$

Both formulae (5.1.2) and (5.1.3) can be proved with the help of the implicit function theorem. This explains the assumptions on the  $\nabla c_i(x)$ 's; things become more delicate when several inequalities are involved to define  $S$ . □

Naturally, the concept of tangency is local, as it depends only on the behaviour of  $S$  near  $x$ . From its definition (5.1.1),  $T_S(x)$  appears as the set of all possible cluster points of the difference quotients  $\{(y - x)/t\}$ , with  $y \in S$  and  $t \downarrow 0$ ; using set-valued notation (see §A.5):

$$T_S(x) = \lim_{t \downarrow 0} \text{ext} \frac{S - x}{t}. \quad (5.1.4)$$

Another interpretation uses the *distance function*  $x \mapsto d_S(x) := \min_{y \in S} \|y - x\|$ :  $T_S(x)$  can also be viewed as the set of  $d$ 's such that

$$\liminf_{t \downarrow 0} \frac{d_S(x + td)}{t} = 0. \quad (5.1.5)$$

Knowing that  $d_S(x) = 0$  when  $x \in S$ , the infimand of (5.1.5) can be interpreted as a difference quotient:  $[d_S(x + td) - d_S(x)]/t$ . Finally, (5.1.5) can be interpreted in a set-formulation: for any  $\varepsilon > 0$  and for any  $\delta > 0$ , there exists  $0 < t \leq \delta$  such that

$$x + td \in S + B(0, t\varepsilon), \quad \text{i.e. } d \in \frac{S - x}{t} + B(0, \varepsilon).$$

**Remark 5.1.5** In (5.1.4), we have taken the tangent cone as a  $\lim \text{ext}$ , which corresponds to a  $\liminf$  in (5.1.5). Another possible approach could have been to define another “tangent cone”, namely

$$\lim_{t \downarrow 0} \text{int} \frac{S - x}{t}.$$

In this case, (5.1.5) would have been changed to

$$\limsup_{t \downarrow 0} \frac{d_S(x + td)}{t} = 0 \quad [= \lim_{t \downarrow 0} d_S(x + td)/t] \quad (5.1.6)$$

(where the second form relies on the fact that  $d_S$  is nonnegative). In a set-formulation as before, we see that (5.1.6) means: for any  $\varepsilon > 0$ , there exists  $\delta > 0$  such that

$$d \in \frac{S - x}{t} + B(0, \varepsilon) \quad \text{for all } 0 < t \leq \delta.$$

We will see in §5.2 below that the pair of alternatives (5.1.5) – (5.1.6) is irrelevant for our purpose, because both definitions coincide when  $S$  is convex.  $\square$

**Remark 5.1.6** Still another “tangent cone” would also be possible: one says that  $d$  is a “feasible direction” for  $S$  at  $x \in S$  when there exists  $\delta > 0$  such that

$$x + td \in S \quad [\text{i.e. } d \in (S - x)/t] \quad \text{for all } 0 < t \leq \delta.$$

Once again, we will see that the difference is of little interest: when  $S$  is convex,  $T_S(x)$  is the closure of the cone of feasible directions thus defined.  $\square$

## 5.2 The Tangent and Normal Cones to a Convex Set

Instead of a general set  $S$ , we now consider a *closed convex* set  $C \subset \mathbb{R}^n$ . In this restricted situation, the tangent cone can be given a more handy expression. The key is to observe that the role of the property  $t_k \downarrow 0$  is special in (5.1.5): when both  $x$  and  $x + t_k d_k$  are in  $C$ , then  $x + \tau d_k \in C$  for all  $\tau \in ]0, t_k]$ . In particular,  $C - \{x\} \subset T_C(x)$ . Indeed, the tangent cone has a *global* character:

**Proposition 5.2.1** *The tangent cone to  $C$  at  $x$  is the closure of the cone generated by  $C - \{x\}$ :*

$$\begin{aligned} T_C(x) &= \overline{\text{cone}}(C - x) = \text{cl } \mathbb{R}^+(C - x) \\ &= \text{cl}\{d \in \mathbb{R}^n : d = \alpha(y - x), y \in C, \alpha \geq 0\}. \end{aligned} \quad (5.2.1)$$

PROOF. We have just said that  $C - \{x\} \subset T_C(x)$ . Because  $T_C(x)$  is a closed cone (Proposition 5.1.3), it immediately follows that  $\text{cl } \mathbb{R}^+(C - x) \subset T_C(x)$ . Conversely, for  $d \in T_C(x)$ , take  $\{x_k\}$  and  $\{t_k\}$  as in the definition (5.1.1): the point  $(x_k - x)/t_k$  is in  $\mathbb{R}^+(C - x)$ , hence its limit  $d$  is in the closure of this latter set.  $\square$

**Remark 5.2.2** This new definition is easier to work with – and to master. Furthermore, it strongly recalls Remark 2.2.4: the term in brackets in (5.2.1) is just a union,

$$\text{cone}(C - x) := \bigcup_{t>0} \frac{C - x}{t}$$

and, thanks to the monotonicity property of the “difference quotient”  $t \mapsto (C - x)/t$ , it is also a limit:

$$\text{cone}(C - x) = \lim_{t \downarrow 0} \frac{C - x}{t}$$

to be compared with the definition (2.2.3) of the asymptotic cone. Having taken a union, or a limit, the closure operation is now necessary, but it was not when we took an intersection. Also, the limit above is unambiguous (it is a union!), and can be understood as the lim ext or the lim int; see Remark 5.1.5. As for Remark 5.1.6, we see that the cone of feasible directions for the convex  $C$  at  $x$  is just the very last set in brackets in (5.2.1).  $\square$

As a closed convex set,  $T_C(x)$  can also be described as an intersection of closed half-spaces – remember §4.2(b). In the present conical situation, some more can be said:

**Definition 5.2.3** The direction  $s \in \mathbb{R}^n$  is said *normal* to  $C$  at  $x \in C$  when

$$\langle s, y - x \rangle \leq 0 \quad \text{for all } y \in C. \quad (5.2.2)$$

The set of all such directions is called *normal cone* to  $C$  at  $x$ , denoted by  $N_C(x)$ .  $\square$

That  $N_C(x)$  is a closed convex cone is clear enough. A normal is a vector  $s$  such that the angle between  $s$  and  $y - x$  is obtuse for all  $y \in C$ . A consequence of §4.2(a) is that there is a nonzero normal at each  $x$  of  $\text{bd } C$ . Indeed, Theorem 3.1.1 tells us that

$$v - p_C(v) \in N_C(p_C(v)) \quad \text{for all } v \in \mathbb{R}^n.$$

By contrast,  $N_C(x) = \{0\}$  for  $x \in \text{int } C$ . As an example, if

$$C = H_{s,r}^- = \{y \in \mathbb{R}^n : \langle s, y \rangle \leq r\}$$

is a closed half-space limited by a given  $H_{s,r}$ , then its normals at any point of  $H_{s,r}$  are the nonnegative multiples of  $s$ .

**Proposition 5.2.4** *The normal cone is the polar of the tangent cone.*

PROOF. If  $\langle s, d \rangle \leq 0$  for all  $d \in C - x$ , the same holds for all  $d \in \mathbb{R}^+(C - x)$ , as well as for all  $d$  in the closure  $T_C(x)$  of the latter. Thus,  $N_C(x) \subset [T_C(x)]^\circ$ .

Conversely, take  $s$  arbitrary in  $[T_C(x)]^\circ$ . The relation  $\langle s, d \rangle \leq 0$ , which holds for all  $d \in T_C(x)$ , a fortiori holds for all  $d \in C - x \subset T_C(x)$ ; this is just (5.2.2).  $\square$

Knowing that the tangent cone is closed, this result can be combined with Proposition 4.2.7 to obtain a third definition:

**Corollary 5.2.5** *The tangent cone is the polar of the normal cone:*

$$T_C(x) = \{d \in \mathbb{R}^n : \langle s, d \rangle \leq 0 \text{ for all } s \in N_C(x)\}.$$

$\square$

This describes  $T_C(x)$  as an intersection of *homogeneous* half-spaces and the relationship with §4.2(b) is clear. With the notation thereof,  $r_s = 0$  for each  $s$ , and the index-set  $\Sigma_{T_C(x)}^0$  is nothing more than  $N_C(x)$ ; see again Remark 4.2.8.

It is interesting to note here that normality is again a local concept, even though (5.2.2) does not suggest it. Indeed the normal cone at  $x$  to  $C \cap B(x, \delta)$  coincides with  $N_C(x)$ . Also, if  $C'$  is “sandwiched”, i.e. if

$$C - \{x\} \subset C' - \{x\} \subset T_C(x),$$

then  $N_{C'}(x) = N_C(x)$  – and  $T_{C'}(x) = T_C(x)$ . Let us add that tangent and normal cones to a nonclosed convex set  $C$  could be defined if needed: just replace  $C$  by  $\text{cl } C$  in the definitions.

Another remark is that the tangent and normal cones are “homogeneous” objects, in that they contain 0 as a distinguished element. It is most often the translated version  $x + T_C(x)$  that is used and visualized; see Fig. 5.1.1 again.

**Examples 5.2.6** (a) If  $C = K$  is a closed convex cone,  $T_K(0) = K$ : the polar  $K^\circ$  of a closed convex cone is its normal cone at 0. On the other hand, if  $0 \neq x \in K$ , then  $T_K(x)$  contains at least one subspace, namely  $\mathbb{R}\{x\}$ . Actually, we even have

$$N_K(x) = \{s \in K^\circ : \langle s, x \rangle = 0\} \quad \text{for } x \neq 0.$$

To see it, observe that  $T_K(x) \supset \{x\}$  means  $N_K(x) \subset \{x\}^\perp$ ; in other words, the relation of definition of  $K^\circ$

$$\langle s, y - x \rangle \leq 0 \quad \text{for all } y \in K$$

reduces to

$$\langle s, y \rangle \leq 0 \quad [= \langle s, x \rangle] \quad \text{for all } y \in K.$$

A cone is a set of vectors defined up to a multiplicative constant, and the value of this constant has often little relevance. As far as the concepts of polarity, tangency, normality are concerned, for example, a closed convex cone  $T$  (or  $N$ ) could equally be replaced by the compact  $T \cap B(0, 1)$ ; or also by  $\{x \in T : \|x\| = 1\}$ , in which the redundancy is totally eliminated.

(b) Take a closed convex polyhedron defined by  $m$  constraints:

$$C := \{x \in \mathbb{R}^n : \langle s_j, x \rangle \leq r_j \text{ for } j = 1, \dots, m\} \quad (5.2.3)$$

and define

$$J(x) := \{j = 1, \dots, m : \langle s_j, x \rangle = r_j\}$$

the set of *active constraints* at  $x \in C$ . Then

$$T_C(x) = \{d \in \mathbb{R}^n : \langle s_j, d \rangle \leq 0 \text{ for } j \in J(x)\},$$

$$N_C(x) = \text{cone}\{s_j : j \in J(x)\} = \left\{ \sum_{j \in J(x)} \alpha_j s_j : \alpha_j \geq 0 \right\}.$$

(c) Let  $C$  be the unit simplex  $\Delta_n$  of Example 1.1.3 and  $\alpha = (\alpha_1, \dots, \alpha_n) \in \Delta_n$ . If each  $\alpha_i$  is positive, i.e. if  $\alpha \in \text{ri } \Delta_n$ , then the tangent cone to  $\Delta_n$  at  $\alpha$  is  $\text{aff } \Delta_n - \{\alpha\}$ , i.e. the linear hyperplane of equation  $\sum_{i=1}^n \alpha_i = 0$ . Otherwise, with  $e := (1, \dots, 1) \in \mathbb{R}^n$ :

$$T_{\Delta_n}(\alpha) = \{d = (d_1, \dots, d_n) : e^\top d = 0, d_i \geq 0 \text{ if } \alpha_i = 0\}.$$

Using Example (b) above, calling  $\{e_1, \dots, e_n\}$  the canonical basis of  $\mathbb{R}^n$  and denoting by  $J(\alpha) := \{j : \alpha_j = 0\}$  the active set at  $\alpha$ , we obtain the normal cone:

$$\begin{aligned} N_{\Delta_n}(\alpha) &= \text{cone}\left[\{e\} \cup \{-e\} \cup_{j \in J(\alpha)} \{-e_j\}\right] \\ &= \left\{ \sum_{j \in \{0\} \cup J(\alpha)} \beta_j e_j : \beta_j \leq 0 \text{ for } j \in J(\alpha) \right\}. \end{aligned}$$

This last example illustrates an interesting *complexity* property: for a closed convex polyhedron described by (5.2.3),

- the tangent cone is conveniently defined as an intersection of (homogeneous) half-spaces: it “resembles”  $C$ ;
- the normal cone is conveniently defined as a conical hull, with a “small” number of generators; its description by closed half-spaces would be tedious.

When characterizing pairs of polar cones, say  $T$  and  $N$ , which are both polyhedral, this kind of duality is usual: one characterization is complex when the other is simple.  $\square$

### 5.3 Some Properties of Tangent and Normal Cones

Let us give some properties of the tangent cone, which result directly from Proposition 5.2.1.

- For fixed  $x$ ,  $T_C(x)$  increases if and only if  $N_C(x)$  decreases, and these properties happen in particular when  $C$  increases.
- The set  $\text{cone}(C - x)$  and its closure  $T_C(x)$  have the same affine hull (actually linear hull!) and the same relative interior. It is not difficult to check that these last sets are  $(\text{aff } C - x)$  and  $\mathbb{R}_*^+(\text{ri } C - x)$  respectively.
- $T_C(x) = \text{aff } C - x$  whenever  $x \in \text{ri } C$  (in particular,  $T_C(x) = \mathbb{R}^n$  if  $x \in \text{int } C$ ). As a result, approximating  $C$  by  $x + T_C(x)$  presents some interest only when  $x \in \text{rbd } C$ . Along this line, we warn the reader against a too sloppy comparison of the tangent cone with the concept of tangency to a surface: with this intuition in mind, one should rather think of the (relative) *boundary* of  $C$  as being approximated by the (relative) boundary of  $T_C(x)$ .
- The concept of tangent cone fits rather well with the convexity-preserving operations of §1.2. Validating the following calculus rules involves elementary arguments only, and is left as an exercise.

**Proposition 5.3.1** *Here, the  $C$ 's are nonempty closed convex sets.*

(i) *For  $x \in C_1 \cap C_2$ , there holds*

$$T_{C_1 \cap C_2}(x) \subset T_{C_1}(x) \cap T_{C_2}(x) \quad \text{and} \quad N_{C_1 \cap C_2}(x) \supset N_{C_1}(x) + N_{C_2}(x).$$

(ii) *With  $C_i \subset \mathbb{R}^{n_i}$ ,  $i = 1, 2$  and  $(x_1, x_2) \in C_1 \times C_2$ ,*

$$T_{C_1 \times C_2}(x_1, x_2) = T_{C_1}(x_1) \times T_{C_2}(x_2),$$

$$N_{C_1 \times C_2}(x_1, x_2) = N_{C_1}(x_1) \times N_{C_2}(x_2).$$

(iii) *With an affine mapping  $A(x) = y_0 + A_0x$  ( $A_0$  linear) and  $x \in C$ ,*

$$T_{A(C)}[A(x)] = \text{cl}[A_0 T_C(x)] \quad \text{and} \quad N_{A(C)}[A(x)] = \bar{A}_0^{-1}[N_C(x)].$$

(iv) *In particular (start from (ii), (iii) and proceed as when proving (1.2.2)):*

$$T_{C_1 + C_2}(x_1 + x_2) = \text{cl}[T_{C_1}(x_1) + T_{C_2}(x_2)],$$

$$N_{C_1 + C_2}(x_1 + x_2) = N_{C_1}(x_1) \cap N_{C_2}(x_2). \quad \square$$

**Remark 5.3.2** To obtain equality in (i), an additional assumption is necessary. One was used in Proposition 2.1.10, see also (2.1.5):

$$0 \in \text{ri}(C_1 - C_2) \quad \text{or} \quad (\text{ri } C_1) \cap (\text{ri } C_2) \neq \emptyset \tag{5.3.1}$$

(the proof of the corresponding statement becomes a bit longer). The gap between the sets of (i) explains many of the technical difficulties that will be encountered later. It also explains that its cure (5.3.1) will also cure these difficulties.  $\square$

Some more properties of tangent and normal cones are worth mentioning, which patch together various notions seen earlier in this chapter.

**Proposition 5.3.3** *For  $x \in C$  and  $s \in \mathbb{R}^n$ , the following properties are equivalent:*

- (i)  $s \in N_C(x)$ ;
- (ii)  $x$  is in the exposed face  $F_C(s)$ :  $\langle s, x \rangle = \max_{y \in C} \langle s, y \rangle$ ;
- (iii)  $x = p_C(x + s)$ .

PROOF. Nothing really new: everything comes from the definitions of normal cones, supporting hyperplanes, exposed faces, and the characteristic property (3.1.3) of the projection operator.  $\square$

This result is illustrated on Fig. 5.3.1 and implies in particular:

$$p_C^{-1}(x) = \{x\} + N_C(x) \quad \text{for all } x \in C.$$

Also,

$$x \neq x' \implies [\{x\} + N_C(x)] \cap [\{x'\} + N_C(x')] = \emptyset$$

(otherwise the projection would not be single-valued).

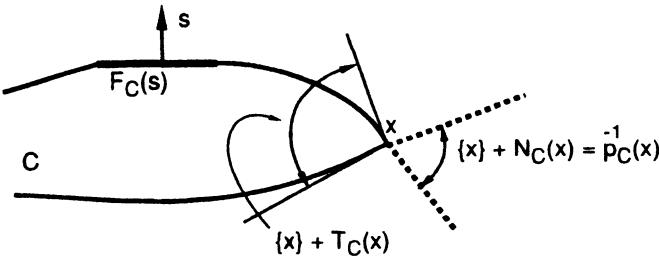


Fig. 5.3.1. Normal cones, projections and exposed faces

**Remark 5.3.4** Let us come back again to Fig. 4.2.2. In a first step, fix  $x \in C$  and consider only those supporting hyperplanes that pass through  $x$ , i.e. those indexed in  $N_C(x)$ . The corresponding intersection of half-spaces just constructs  $T_C(x)$  and ends up with

$$\{x\} + T_C(x) \supset C.$$

Note in passing that the closure operation of Proposition 5.2.1 is necessary when  $\text{rbd } C$  presents some curvature near  $x$ .

Then, in a second step, do this operation for all  $x \in C$ :

$$C \subset \bigcap_{x \in C} [x + T_C(x)]. \quad (5.3.2)$$

A first observation is that  $x$  can actually be restricted to the relative boundary of  $C$ : for  $x \in \text{ri } C$ ,  $T_C(x)$  expands to the whole  $\text{aff } C - x$  and contains all other tangent cones. A second observation is that (5.3.2) actually holds as an equality. In fact, write a point  $y \notin C$

as  $y = p_C(y) + s$  with  $s = y - p_C(y)$ ; Proposition 5.3.3 tells us that  $s \in N_C[p_C(y)]$ , hence the nonzero  $s$  cannot be in  $T_C[p_C(y)]$ : we have established  $y \notin p_C(y) + T_C[p_C(y)]$ ,  $y$  is not in the right-hand side of (5.3.2). In a word:

$$C = \bigcap_{x \in \text{rbd } C} [x + T_C(x)],$$

which sheds a new light on the outer description of  $C$  discussed in §4.2(b).  $\square$

We conclude with an interesting approximation result. As indicated by (3.1.6), the projection onto a fixed convex set is a continuous operator. More can actually be said, and the normal and tangent cones help estimating the variation of this projection. The following result can be proved using the material developed in this chapter.

**Proposition 5.3.5** *For given  $x \in C$  and  $d \in \mathbb{R}^n$ , there holds*

$$\lim_{t \downarrow 0} \frac{p_C(x + td) - x}{t} = p_{T_C(x)}(d). \quad (5.3.3)$$

HINT. Start from the characterization (3.1.3) of a projection, to observe that the difference quotient  $[p_C(x + td) - x]/t$  is the projection of  $d$  onto  $(C - x)/t$ . Then let  $t \downarrow 0$ ; the result comes with the help of (5.1.4) and Remark 5.2.2.  $\square$

This result is illustrated on Fig. 5.3.2. It gives us a sort of derivative, more precisely the *directional derivative* of  $p_C$  at  $x$ . If  $x \in \text{ri } C$ , then  $T_C(x)$  becomes the subspace parallel to  $\text{aff } C$  and we recover the linearity (at least local) of the projection operator.

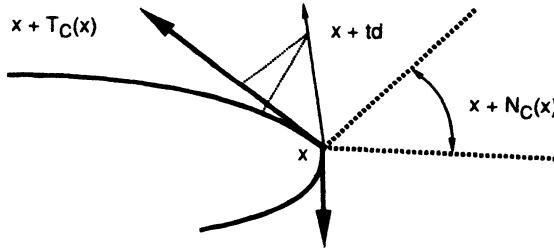


Fig. 5.3.2. Differentiating a projection

In view of (5.3.3), the notation

$$p_{T_C(x)}(\cdot) = p'_C(x, \cdot)$$

is natural (it has already been seen in §I.4.1 in a one-dimensional setting). Moreover, remembering the approximation role of tangent cones, a possible (although daring) notation is also  $T_C(x) = C'(x)$ ; then (5.3.3) can be rephrased as: the projection and derivation operations commute:  $p_{C'(x)}(\cdot) = p'_C(x, \cdot)$ .

## IV. Convex Functions of Several Variables

**Prerequisites.** Basic definitions and properties of convex sets (Chap. III); basic results on the analysis of functions of several variables; and to support intuition if necessary: convex functions of one real variable (Chap. I).

**Introduction.** The study of convex functions goes together with that of convex sets; accordingly, this chapter and the previous one constitute the first serious steps into the world of convex analysis. Most of the concepts to come have already been seen in Chap. I; a reader mastering that chapter should therefore have no major difficulty following our development. Nevertheless, some of the definitions and properties introduced or proved in a simple one-dimensional setting may become harder to visualize when several variables come into play: the natural ordering of  $\mathbb{R}$  is no longer present to help.

This chapter has no pretension to exhaustivity; similarly to Chap. III, it has been kept minimal, containing what is necessary to comprehend the sequel. Furthermore, it contains many examples commonly appearing in convex optimization, like: piecewise affine and quadratic functions, max-functions, functions associated to convex sets (indicator, support, distance functions).

### 1 Basic Definitions and Examples

#### 1.1 The Definitions of a Convex Function

**Definition 1.1.1** Let  $C$  be a nonempty convex set in  $\mathbb{R}^n$ . A function  $f : C \rightarrow \mathbb{R}$  is said to be *convex* on  $C$  when, for all pairs  $(x, x') \in C \times C$  and all  $\alpha \in ]0, 1[$ , there holds

$$f(\alpha x + (1 - \alpha)x') \leq \alpha f(x) + (1 - \alpha)f(x'). \quad (1.1.1)$$

□

We say that  $f$  is *strictly convex* on  $C$  when (1.1.1) holds as a strict inequality if  $x \neq x'$ . An even stronger property is that there exists  $c > 0$  such that

$$f(\alpha x + (1 - \alpha)x') \leq \alpha f(x) + (1 - \alpha)f(x') - \frac{1}{2}c\alpha(1 - \alpha)\|x - x'\|^2 \quad (1.1.2)$$

for all  $(x, x') \in C \times C$  and all  $\alpha \in ]0, 1[$ . In this case,  $f$  is said to be *strongly convex* on  $C$  (with *modulus* of strong convexity  $c$ ). Passing from (1.1.1) to (1.1.2) does not change much the class of functions considered:

**Proposition 1.1.2** *The function  $f$  is strongly convex on  $C$  with modulus  $c$  if and only if the function  $f - \frac{1}{2}c\|\cdot\|^2$  is convex on  $C$ .*

PROOF. Use direct calculations in the definition (1.1.1) of convexity applied to the function  $f - \frac{1}{2}c\|\cdot\|^2$ , namely:

$$\begin{aligned} f(\alpha x + (1 - \alpha)x') - \frac{1}{2}c\|\alpha x + (1 - \alpha)x'\|^2 &\leqslant \\ &\leqslant \alpha f(x) + (1 - \alpha)f(x') - \frac{1}{2}c[\alpha\|x\|^2 + (1 - \alpha)\|x'\|^2]. \end{aligned}$$
□

Although simple, this statement illustrates a useful technique in convex analysis: to prove that a convex function has a certain property, one establishes a related property on a suitable strongly convex perturbation of the given function.

The set  $C$  needed in Definition 1.1.1 (which can be the whole space) appears as a sort of domain of definition of  $f$ . Of course, it has to be convex so that the left-hand side of (1.1.1) makes sense. In a more modern definition, a convex function  $f$  is considered as defined on the whole of  $\mathbb{R}^n$ , but possibly taking infinite values:

**Definition 1.1.3 (The Set  $\text{Conv } \mathbb{R}^n$ )** A function  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ , not identically  $+\infty$ , is said to be convex when, for all  $(x, x') \in \mathbb{R}^n \times \mathbb{R}^n$  and all  $\alpha \in ]0, 1[$ , there holds

$$f(\alpha x + (1 - \alpha)x') \leqslant \alpha f(x) + (1 - \alpha)f(x'),$$

considered as an inequality in  $\mathbb{R} \cup \{+\infty\}$ .

The class of such functions is denoted by  $\text{Conv } \mathbb{R}^n$ .

□

We mention here that our definition coincides with that of *proper* convexity used by other authors. The distinction is necessary when the value  $f(x) = -\infty$  is allowed; but this value is excluded from the very beginning in the present book.

To realize the equivalence between our two definitions, extend an  $f$  from Definition 1.1.1 by

$$f(x) := +\infty \quad \text{for } x \notin C, \tag{1.1.3}$$

thus obtaining a new  $f$ , which is now in  $\text{Conv } \mathbb{R}^n$ . Conversely, consider the following definition (meaningful even for nonconvex  $f$ , incidentally):

**Definition 1.1.4** The *domain* (or effective domain) of  $f \in \text{Conv } \mathbb{R}^n$  is the nonempty set

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

Clearly enough, an  $f$  satisfying (1.1.1), (1.1.3) has a convex domain; given  $f \in \text{Conv } \mathbb{R}^n$ , we can therefore take  $C := \text{dom } f$  to obtain a convex function in the sense of Definition 1.1.1. Strong convexity is also defined in the spirit of Definition 1.1.3, via (1.1.2) with  $x$  and  $x'$  varying in  $\text{dom } f$  or in  $\mathbb{R}^n$ : it makes no difference. Same remark for strict convexity (checking all these claims is a good exercise to familiarize oneself with computations in  $\mathbb{R} \cup \{+\infty\}$ ).

Now, we recall that the *graph* of an arbitrary function is the set of couples  $(x, f(x))$  in  $\mathbb{R}^n \times \mathbb{R}$ . When moving to the unilateral world of convex analysis, the following is relevant:

**Definition 1.1.5** Given  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ , not identically equal to  $+\infty$ , the *epigraph* of  $f$  is the nonempty set

$$\text{epi } f := \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : r \geq f(x)\}.$$

Its *strict epigraph*  $\text{epi}_s f$  is defined likewise, with “ $\geq$ ” replaced by “ $>$ ” (beware that the word “strict” here has nothing to do with strict convexity).  $\square$

In terms of *sublevel-sets*, we have the equivalent definition

$$(x, r) \in \text{epi } f \iff x \in S_r(f) [= \{x \in \mathbb{R}^n : f(x) \leq r\}]. \quad (1.1.4)$$

The following property is easy to derive, and can be interpreted as giving one more definition of convex functions, which is now of *geometric* nature.

**Proposition 1.1.6** Let  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  be not identically equal to  $+\infty$ . The three properties below are equivalent:

- (i)  $f$  is convex in the sense of Definition 1.1.3;
- (ii) its epigraph is a convex set in  $\mathbb{R}^n \times \mathbb{R}$ ;
- (iii) its strict epigraph is a convex set in  $\mathbb{R}^n \times \mathbb{R}$ .

PROOF. Left as an exercise.  $\square$

We say that  $f$  is *concave* when  $-f$  is convex, or equivalently when the *hypograph* of  $f$  (revert the inequality in Definition 1.1.5) is a convex set. We will see on examples that either the analytical Definition 1.1.3 or the geometric one coming from 1.1.6 may be more convenient, depending on the situation.

**Remark 1.1.7** The sublevel-sets of  $f \in \text{Conv } \mathbb{R}^n$  are convex (possibly empty) subsets of  $\mathbb{R}^n$ . To construct  $S_r(f)$ , we cut the epigraph of  $f$  by a horizontal blade, forming the intersection  $\text{epi } f \cap (\mathbb{R}^n \times \{r\})$  of two convex sets; then we project the result down to  $\mathbb{R}^n \times \{0\}$  and we change the environment space from  $\mathbb{R}^n \times \mathbb{R}$  to  $\mathbb{R}^n$ . Even though this latter operation changes the topology, it changes neither the closure nor the relative interior.

Conversely, a function whose sublevel-sets are all convex *need not* be convex (see Fig. 1.1.1); such a function is called *quasi-convex*.  $\square$

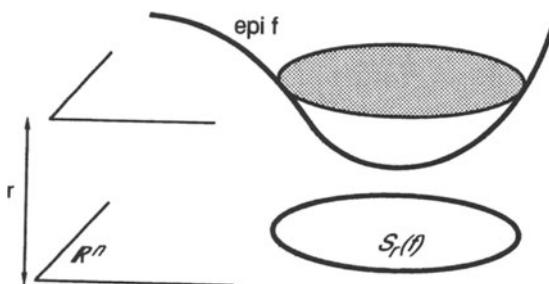


Fig. 1.1.1. Forming a sublevel-set

Observe that  $\text{dom } f$  is the union of the sublevel-sets  $S_r(f)$ , which form a nested family; it is also the projection of  $\text{epi } f \subset \mathbb{R}^n \times \mathbb{R}$  onto  $\mathbb{R}^n$  (so, Proposition III.1.2.4 confirms its convexity). We will see later that a convex function behaves nicely on the interior of its domain, and even on its relative interior. By contrast, anything can happen on the (relative) boundary of  $\text{dom } f$ ; of course,  $f$  can be infinite there, but can also behave much more strangely than in the univariate case.

The basic inequality (1.1.1) can be generalized to convex combinations of more than two points:

**Theorem 1.1.8** *Let  $f \in \text{Conv } \mathbb{R}^n$ . Then, for all collections  $\{x_1, \dots, x_k\}$  of points in  $\text{dom } f$  and all  $\alpha = (\alpha_1, \dots, \alpha_k)$  in the unit simplex of  $\mathbb{R}^k$ , there holds the inequality of Jensen (in summation form)*

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

PROOF. Nothing particular: just proceed as in Theorem I.1.2.1.  $\square$

Starting from  $f \in \text{Conv } \mathbb{R}^n$ , we thus construct the convex set  $\text{epi } f$ . Conversely, if  $E \subset \mathbb{R}^n \times \mathbb{R}$  is the epigraph of a function in  $\text{Conv } \mathbb{R}^n$ , this function is directly obtained from

$$f(x) = \inf \{r : (x, r) \in E\}$$

(recall that  $\inf \emptyset = +\infty$ ; we will see in §1.3(g) what sets are epigraphs of a convex function). In view of this correspondence, the properties of a convex function  $f$  are intimately related to those developed in Chap. III, applied to  $\text{epi } f$ . For example, we will see later that important functions, maybe the most important in optimization, are those having a closed epigraph. Also, it is clear that  $\text{aff epi } f$  contains the vertical lines  $\{x\} \times \mathbb{R}$ , with  $x \in \text{dom } f$ . This shows that  $\text{epi } f$  cannot be an open set, nor relatively open: take points of the form  $(x, f(x) - \varepsilon)$ . As a result,  $\text{ri epi } f$  cannot be an epigraph, but it is nevertheless of interest to see how this set is constructed:

**Proposition 1.1.9** *Let  $f \in \text{Conv } \mathbb{R}^n$ . The relative interior of  $\text{epi } f$  is the union over  $x \in \text{ri dom } f$  of the open non-majorized intervals with bottom endpoints at  $f(x)$ :*

$$\text{ri epi } f = \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : x \in \text{ri dom } f, r > f(x)\}.$$

PROOF. Since  $\text{dom } f$  is the image of  $\text{epi } f$  under the linear mapping ‘‘projection onto  $\mathbb{R}^n$ ’’, Propositions III.1.2.4 and III.2.1.12 tell us that

$$\text{ri dom } f \text{ is the projection onto } \mathbb{R}^n \text{ of } \text{ri epi } f. \quad (1.1.5)$$

Now take  $x$  arbitrary in  $\text{ri dom } f$ . The subset of  $\text{ri epi } f$  that is projected onto  $x$  is just  $(\{x\} \times \mathbb{R}) \cap \text{ri epi } f$ , which in turn is  $\text{ri}[(\{x\} \times \mathbb{R}) \cap \text{epi } f]$  (use Proposition III.2.1.10). This latter set is clearly  $]f(x), +\infty[$ .

In summary, we have proved that, for  $x \in \text{ri dom } f$ ,  $(x, r) \in \text{ri epi } f$  if and only if  $r > f(x)$ . Together with (1.1.5), this proves our claim.  $\square$

Beware that  $\text{ri epi } f$  is not the strict epigraph of  $f$  (watch the side-effect on the relative boundary of  $\text{dom } f$ ).

## 1.2 Special Convex Functions: Affinity and Closedness

In view of their Definition 1.1.6(ii), convex functions can be classified on the basis of a classification of convex sets in  $\mathbb{R}^n \times \mathbb{R}$ .

**(a) Linear and Affine Functions** The epigraph of a linear function is characterized by  $s \in \mathbb{R}^n$ , and is made of those  $(x, r) \in \mathbb{R}^n \times \mathbb{R}$  such that  $r \geq \langle s, x \rangle$ .

Next, we find the epigraphs of *affine functions*  $f$ , which are conveniently written in terms of some  $x_0 \in \mathbb{R}^n$ :

$$\{(x, r) : r \geq f(x_0) + \langle s, x - x_0 \rangle\} = \{(x, r) : \langle s, x \rangle - r \leq \langle s, x_0 \rangle - f(x_0)\}.$$

In the language of convex sets, the epigraph of an affine function is a closed half-space, characterized by (a constant term and) a vector  $(s, -1) \in \mathbb{R}^n \times \mathbb{R}$ ; the essential property of this vector is to be *non-horizontal*. Affine functions thus play a special role, just as half-spaces did in Chap. III. This explains the interest of the next result; it says a little more than Lemma III.4.2.1, and is actually of paramount importance.

**Proposition 1.2.1** *Any  $f \in \text{Conv } \mathbb{R}^n$  is minorized by some affine function. More precisely: for any  $x_0 \in \text{ri dom } f$ , there is  $s$  in the subspace parallel to  $\text{aff dom } f$  such that*

$$f(x) \geq f(x_0) + \langle s, x - x_0 \rangle \quad \text{for all } x \in \mathbb{R}^n.$$

*In other words, the affine function can be forced to coincide with  $f$  at  $x_0$ .*

PROOF. We know that  $\text{dom } f$  is the image of  $\text{epi } f$  under the linear mapping “projection onto  $\mathbb{R}^n$ ”. Look again at the definition of an affine hull (§III.1.3) to realize that

$$\text{aff}(\text{epi } f) = \text{aff}(\text{dom } f) \times \mathbb{R}.$$

Denote by  $V$  the linear subspace parallel to  $\text{aff}(\text{dom } f)$ , so that  $\text{aff}(\text{dom } f) = \{x_0\} + V$  with  $x_0$  arbitrary in  $\text{dom } f$ ; then we have

$$\text{aff}(\text{epi } f) = \{x_0 + V\} \times \mathbb{R}. \tag{1.2.1}$$

We equip  $V \times \mathbb{R}$  and  $\mathbb{R}^n \times \mathbb{R}$  with the scalar product of product-spaces.

Choose  $x_0 \in \text{ri dom } f$ . Then Proposition 1.1.9 tells us that  $(x_0, f(x_0)) \in \text{rbd epi } f$  and we can take a nontrivial hyperplane supporting  $\text{epi } f$  at  $(x_0, f(x_0))$ : using Remark III.4.2.2 and (1.2.1), there are  $s = s_V \in V$  and  $\alpha \in \mathbb{R}$ , not both zero, such that

$$\langle s, x \rangle + \alpha r \leq \langle s, x_0 \rangle + \alpha f(x_0) \tag{1.2.2}$$

for all  $(x, r)$  with  $f(x) \leq r$ .

Because of our choice of  $s$  (in  $V$ ) and  $x_0$  (in  $\text{ri dom } f$ ), we can take  $\delta > 0$  so small that  $x_0 + \delta s \in \text{dom } f$ , for which (1.2.2) gives

$$\delta \|s\|^2 \leq \alpha [f(x_0) - f(x_0 + \delta s)] < +\infty;$$

this shows  $\alpha \neq 0$  (otherwise, both  $s$  and  $\alpha$  would be zero). Without loss of generality, we can assume  $\alpha = -1$ ; then (1.2.2) gives our affine function.  $\square$

Once again, the importance of this result cannot be over-emphasized. With respect to Lemma III.4.2.1, it says that a convex epigraph is supported by a *non-vertical* hyperplane. Among its consequences, we see that a convex function, having an affine minorant, is bounded from below on every bounded set of  $\mathbb{R}^n$ .

As already said on several occasions, important special convex sets are the cones, and conical epigraphs will deserve the full Chap.V. So we turn to another class, especially important for optimization, namely closed convex sets.

**(b) Closed Convex Functions** For a (convex) function to have a minimum, a very first requirement is *lower semi-continuity* (l.s.c.). This property is therefore of fundamental importance for our subsequent developments: if we want to minimize a function  $f$  on some compact set  $K$ , we do not need to bother with existence if  $f$  is known to be lower semi-continuous on  $\mathbb{R}^n$ ; and this holds even if  $K$  is not contained in  $\text{dom } f$  – an appreciable formal advantage.

First, we give some material, independently of any convexity. A function  $f$  is lower semi-continuous if, for each  $x \in \mathbb{R}^n$ ,

$$\liminf_{y \rightarrow x} f(y) \geq f(x). \quad (1.2.3)$$

This relation has to hold in  $\mathbb{R} \cup \{+\infty\}$ , which complicates things a little; so the following geometric characterizations are useful:

**Proposition 1.2.2** *For  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ , the following three properties are equivalent:*

- (i)  *$f$  is lower semi-continuous on  $\mathbb{R}^n$ ;*
- (ii)  *$\text{epi } f$  is a closed set in  $\mathbb{R}^n \times \mathbb{R}$ ;*
- (iii) *the sublevel-sets  $S_r(f)$  are closed (possibly empty) for all  $r \in \mathbb{R}$ .*

PROOF.  $[(i) \Rightarrow (ii)]$  Let  $\{(y_k, r_k)\}$  be a sequence of  $\text{epi } f$  converging to  $(x, r)$  for  $k \rightarrow +\infty$ . Since  $f(y_k) \leq r_k$  for all  $k$ , the l.s.c. relation (1.2.3) readily gives

$$r = \lim r_k \geq \liminf_{y \rightarrow x} f(y) \geq \liminf_{y \rightarrow x} f(y) \geq f(x),$$

i.e.  $(x, r) \in \text{epi } f$ .

$[(ii) \Rightarrow (iii)]$  Construct the sublevel-sets  $S_r(f)$  as in Remark 1.1.7: the closed sets  $\text{epi } f$  and  $\mathbb{R}^n \times \{r\}$  have a closed intersection.

$[(iii) \Rightarrow (i)]$  Suppose  $f$  is not lower semi-continuous at some  $x$ : there is a (sub)sequence  $\{y_k\}$  converging to  $x$  such that  $f(y_k)$  converges to  $\rho < f(x) \leq +\infty$ . Pick  $r \in ]\rho, f(x)[$ : for  $k$  large enough,  $f(y_k) \leq r < f(x)$ ; hence  $S_r(f)$  contains the tail of  $\{y_k\}$  but not its limit  $x$ . Consequently, this  $S_r(f)$  is not closed.  $\square$

Beware that, with Definition 1.1.1 in mind, the above statement (i) means more than lower semi-continuity of the restriction of  $f$  to  $C$ : in (1.2.3),  $x$  need not be in  $\text{dom } f$ . Note also that these concepts and results are independent from convexity. Thus, we are entitled to consider the following definition:

**Definition 1.2.3** The function  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  is said to be *closed* if it is lower semi-continuous everywhere, or if its epigraph is closed, or if its sublevel-sets are closed.  $\square$

The next step is to take the lower semi-continuous hull of a function  $f$ , whose value at  $x \in \mathbb{R}^n$  is  $\liminf_{y \rightarrow x} f(y)$ . In view of the proof of Proposition 1.2.2, this operation amounts to closing  $\text{epi } f$ . When doing so, however, we may slide down to  $-\infty$ .

**Definition 1.2.4** The *closure* (or lower semi-continuous hull) of a function  $f$  is the function  $\text{cl } f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{\pm\infty\}$  defined by:

$$\text{cl } f(x) := \liminf_{y \rightarrow x} f(y) \quad \text{for all } x \in \mathbb{R}^n, \quad (1.2.4)$$

or equivalently

$$\text{epi}(\text{cl } f) := \text{cl}(\text{epi } f). \quad (1.2.5) \quad \square$$

An l.s.c. hull may be fairly complicated to compute, though; furthermore, the gap between  $f(y)$  and  $\text{cl } f(y)$  may be impossible to control when  $y$  varies in a neighborhood of a given point  $x$ . Now convexity enters into play and makes things substantially easier, without additional assumption on  $f$  in the above definition:

- First of all, a convex function is minorized by an affine function (Proposition 1.2.1); closing it cannot introduce the value  $-\infty$ .
- Second, the issue reduces to the one-dimensional setting, thanks to the following *radial construction* of  $\text{cl } f$ .

**Proposition 1.2.5** Let  $f \in \text{Conv } \mathbb{R}^n$  and  $x' \in \text{ri dom } f$ . There holds (in  $\mathbb{R} \cup \{+\infty\}$ )

$$(\text{cl } f)(x) = \lim_{t \downarrow 0} f(x + t(x' - x)) \quad \text{for all } x \in \mathbb{R}^n. \quad (1.2.6)$$

PROOF. Since  $x_t := x + t(x' - x) \rightarrow x$  when  $t \downarrow 0$ , we certainly have

$$(\text{cl } f)(x) \leq \liminf_{t \downarrow 0} f(x + t(x' - x)).$$

We will prove the converse inequality by showing that

$$\limsup_{t \downarrow 0} f(x + t(x' - x)) \leq r \quad \text{for all } r \geq (\text{cl } f)(x)$$

(non-existence of such an  $r$  means that  $\text{cl } f(x) = +\infty$ , the proof is finished).

Thus let  $(x, r) \in \text{epi}(\text{cl } f) = \text{cl}(\text{epi } f)$ . Pick  $r' > f(x')$ , hence  $(x', r') \in \text{ri epi } f$  (Proposition 1.1.9). Applying Lemma III.2.1.6 to the convex set  $\text{epi } f$ , we see that

$$t(x', r') + (1-t)(x, r) \in \text{ri epi } f \subset \text{epi } f \quad \text{for all } t \in ]0, 1].$$

This just means

$$f(x + t(x' - x)) \leqslant tr' + (1-t)r \quad \text{for all } t \in ]0, 1]$$

and our required inequality follows by letting  $t \downarrow 0$ .  $\square$

Another way of expressing the same thing is that, to compute  $\text{cl } f$  at some point  $x$ , it suffices to consider the restriction of  $f$  to a half-line, say  $x + \mathbb{R}^+ d$ , meeting  $\text{ri dom } f$ ; here,  $d$  stands for  $x' - x$ . The resulting one-dimensional function  $\varphi(t) := \text{cl } f(x + td)$  becomes “continuous” from the right at  $t = 0$ , in the sense that  $\varphi(0) = \lim_{t \downarrow 0} \varphi(t)$  – an equality in  $\mathbb{R} \cup \{+\infty\}$ .

Some simple but important properties come in conjunction with the results of the previous chapters:

**Proposition 1.2.6** *For  $f \in \text{Conv } \mathbb{R}^n$ , there holds*

$$\text{cl } f \in \text{Conv } \mathbb{R}^n; \quad (1.2.7)$$

$$\text{cl } f \text{ and } f \text{ coincide on the relative interior of } \text{dom } f. \quad (1.2.8)$$

PROOF. We already know from Proposition III.1.2.7 that  $\text{epi cl } f = \text{cl epi } f$  is a convex set; also  $\text{cl } f \leqslant f \not\equiv +\infty$ ; finally, Proposition 1.2.1 guarantees in the relation of definition (1.2.4) that  $\text{cl } f(x) > -\infty$  for all  $x$ : (1.2.7) does hold.

On the other hand, suppose  $x \in \text{ri dom } f$ ; then, the one-dimensional function  $\varphi(t) = f(x + td)$  is continuous at  $t = 0$  (Theorem I.3.1.1); it follows that  $\text{cl } f$  coincides with  $f$  on  $\text{ri dom } f$ ; besides,  $\text{cl } f(x)$  is obviously equal to  $f(x) = +\infty$  for all  $x \notin \text{cl dom } f$ . Altogether, (1.2.8) is true.  $\square$

In particular, a finite-valued convex function ( $\text{dom } f = \mathbb{R}^n$ ) is lower semi-continuous; actually, Theorem 3.1.2 below will confirm that it is more than that: it is continuous, and even locally Lipschitzian.

Due to their importance, closed convex functions deserve a special notation:

**Notation 1.2.7 (The Set  $\overline{\text{Conv}} \mathbb{R}^n$ )** The set of closed convex functions on  $\mathbb{R}^n$  is denoted by  $\overline{\text{Conv}} \mathbb{R}^n$ .  $\square$

**(c) Outer Construction of Closed Convex Functions** The property proved in Proposition 1.2.5 corresponds to a *direct* (or inner) construction of  $\text{cl } f$  from (1.2.4). Equivalently,  $\text{cl } f$  can be constructed as the largest l.s.c. (convex) function minorizing  $f$ . Correspondingly, the closed (convex) set  $\text{epi cl } f$  can also be described *externally*, as an intersection of closed (convex) sets. In view of §III.4.2(b), these closed convex sets can be restricted to be closed half-spaces: convexity provides another simplification of the closure operation. Besides, in view of Proposition 1.2.1, these half-spaces can be assumed non-vertical.

**Proposition 1.2.8** *The closure of  $f \in \text{Conv } \mathbb{R}^n$  is the supremum of all affine functions minorizing  $f$ :*

$$\text{cl } f(x) = \sup_{(s,b) \in \mathbb{R}^n \times \mathbb{R}} \{ \langle s, x \rangle - b : \langle s, y \rangle - b \leqslant f(y) \text{ for all } y \in \mathbb{R}^n \}. \quad (1.2.9)$$

PROOF. A closed half-space containing  $\text{epi } f$  is characterized by a nonzero vector  $(s, \alpha) \in \mathbb{R}^n \times \mathbb{R}$  and a real number  $b$  such that

$$\langle s, x \rangle + \alpha r \leq b \quad \text{for all } (x, r) \in \text{epi } f \quad (1.2.10)$$

(we equip the graph-space  $\mathbb{R}^n \times \mathbb{R}$  with the scalar product of a product-space). Let us denote by  $\Sigma \subset \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}$  the index-set of such triples  $\sigma = (s, \alpha, b)$  and by

$$H_\sigma^- := \{(x, r) : \langle s, x \rangle + \alpha r \leq b\} \quad (1.2.11)$$

the corresponding half-spaces. In other words,

$$\text{epi}(\text{cl } f) = \text{cl}(\text{epi } f) = \cap_{\sigma \in \Sigma} H_\sigma^-.$$

Because of the particular nature of an epigraph, (1.2.10) implies  $\alpha \leq 0$  and, by positive homogeneity, the values  $\alpha = 0$  and  $\alpha = -1$  suffice:  $\Sigma$  can be partitioned in

$$\Sigma_1 := \{(s, -1, b) : (1.2.10) \text{ holds with } \alpha = -1\}$$

and

$$\Sigma_0 := \{(s, 0, b) : (1.2.10) \text{ holds with } \alpha = 0\}.$$

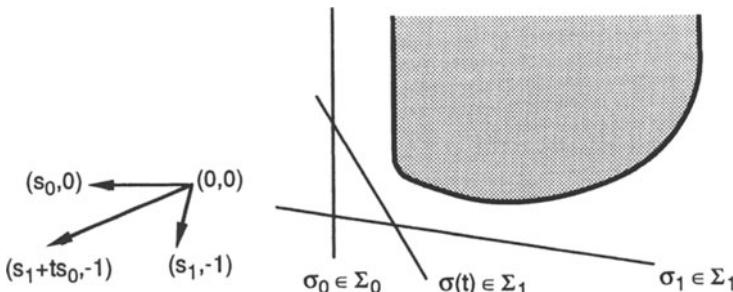
Indeed,  $\Sigma_1$  corresponds to affine functions minorizing  $f$  (Proposition 1.2.1 tells us that  $\Sigma_1 \neq \emptyset$ ) and  $\Sigma_0$  to closed half-spaces of  $\mathbb{R}^n$  containing  $\text{dom } f$  (note that  $\Sigma_0 = \emptyset$  if  $\text{dom } f = \mathbb{R}^n$ ).

We have to prove that, even when  $\Sigma_0 \neq \emptyset$ , intersecting the half-spaces  $H_\sigma^-$  over  $\Sigma$  or over  $\Sigma_1$  produces the same set, namely  $\text{cl epi } f$ . For this we take arbitrary  $\sigma_0 = (s_0, 0, b_0) \in \Sigma_0$  and  $\sigma_1 = (s_1, -1, b_1) \in \Sigma_1$ , we set

$$\sigma(t) := (s_1 + ts_0, -1, b_1 + tb_0) \in \Sigma_1 \quad \text{for all } t \geq 0,$$

and we prove (see Fig. 1.2.1)

$$H_{\sigma_0}^- \cap H_{\sigma_1}^- = \cap_{t \geq 0} H_{\sigma(t)}^- =: H^-.$$



**Fig. 1.2.1.** Closing a convex epigraph

It results directly from the definition (1.2.11) that an  $(x, r)$  in  $H_{\sigma_0}^- \cap H_{\sigma_1}^-$  satisfies

$$\langle s_1 + ts_0, x \rangle - (b_1 + tb_0) \leq r \quad \text{for all } t \geq 0, \quad (1.2.12)$$

i.e.  $(x, r) \in H^-$ . Conversely, take  $(x, r) \in H^-$ . Set  $t = 0$  in (1.2.12) to see that  $(x, r) \in H_{\sigma_1}^-$ . Also, divide by  $t > 0$  and let  $t \rightarrow +\infty$  to see that  $(x, r) \in H_{\sigma_0}^-$ . The proof is complete.  $\square$

### 1.3 First Examples

**(a) Indicator and Support Functions** Given a nonempty subset  $S \subset \mathbb{R}^n$ , the function  $I_S : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  defined by

$$I_S(x) := \begin{cases} 0 & \text{if } x \in S, \\ +\infty & \text{if not} \end{cases}$$

is called the *indicator* function of  $S$ . We mention here that other notations commonly encountered in the literature are  $\delta_S$ ,  $\psi_S$ , or even  $\chi_S$ . Clearly enough,  $I_S$  is [closed and] convex if and only if  $S$  is [closed and] convex. Indeed,  $\text{epi } I_S = S \times \mathbb{R}^+$  by definition.

More generally, if  $f \in \text{Conv } \mathbb{R}^n$  and if  $C$  is a nonempty convex set, the function

$$\varphi(x) := \begin{cases} f(x) & \text{if } x \in C, \\ +\infty & \text{if not} \end{cases}$$

is again convex under one condition: that  $\text{dom } f$  and  $C$  have a nonempty intersection (otherwise  $\varphi$  would be identically  $+\infty$ ). Furthermore,  $\varphi$  is closed when so are  $f$  and  $C$ . Observe in passing that  $\varphi = f + I_C$ .

Attached to a nonempty subset  $S$ , another function of interest is the *support* function of  $S$ , already encountered in Remark III.4.1.2:

$$\sigma_S(x) := \sup \{\langle s, x \rangle : s \in S\}.$$

It turns out to be closed and convex; this is already suggested by Proposition 1.2.8 and will be confirmed below in §2.1(b). Actually, the importance of this function will motivate an extensive development in Chap. V. Here, we just observe that, for  $\alpha > 0$ ,

$$\sup_{s \in S} \langle s, \alpha x \rangle = \alpha \sup_{s \in S} \langle s, x \rangle,$$

hence  $\sigma_S(\alpha x) = \alpha \sigma_S(x)$ : the epigraph of a support function is not only closed and convex, but it is a cone in  $\mathbb{R}^n \times \mathbb{R}$ . Its domain is also a convex cone in  $\mathbb{R}^n$ :

$$\text{dom } \sigma_S = \{a \in \mathbb{R}^n : \exists r \text{ such that } \langle s, a \rangle \leq r \text{ for all } s \in S\}.$$

**(b) Piecewise Affine and Polyhedral Functions** Let  $(s_1, b_1), \dots, (s_m, b_m)$  be  $m$  elements of  $\mathbb{R}^n \times \mathbb{R}$  and consider the function

$$\mathbb{R}^n \ni x \mapsto \check{f}(x) := \max \{ \langle s_j, x \rangle - b_j : j = 1, \dots, m \}. \quad (1.3.1)$$

Such a function is suggestively called *piecewise affine*:  $\mathbb{R}^n$  is divided into (at most  $m$ ) regions in which  $\check{f}$  is affine: the  $j_0^{\text{th}}$  region, possibly empty, is the closed convex polyhedron

$$\{x \in \mathbb{R}^n : \langle s_{j_0}, x \rangle - b_{j_0} \geq \langle s_j, x \rangle - b_j \text{ for } j = 1, \dots, m\}.$$

This terminology is slightly ambiguous, though: a function whose graph is made up of pieces of affine hyperplanes need not be convex, while (1.3.1) can be seen to produce convex functions only (just as with a support function, convexity and closedness of  $\check{f}$  will be confirmed below). It can even be seen that  $\text{epi } \check{f}$  is a closed convex polyhedron; but again, (1.3.1) cannot describe all polyhedral epigraphs.

A *polyhedral* function will be a function whose epigraph is a closed convex polyhedron. Its most general form is given by Definition III.4.2.6:

$$\text{epi } f = \{(x, r) \in \mathbb{R}^n \times \mathbb{R} : \langle s_j, x \rangle + \alpha_j r \leq b_j \text{ for } j \in J\},$$

where  $J$  is a finite set, the  $(s, \alpha, b)_j$  being given in  $\mathbb{R}^n \times \mathbb{R} \times \mathbb{R}$ ,  $(s_j, \alpha_j) \neq 0$  (and  $\mathbb{R}^n \times \mathbb{R}$  is equipped with the scalar product of a product-space). For this set to be an epigraph, each  $\alpha_j$  must be nonpositive and, if  $\alpha_j < 0$ , we may assume without loss of generality  $\alpha_j = -1$ . Furthermore, we may denote by  $\{1, \dots, m\}$  the subset of  $J$  such that  $\alpha_j = -1$ , and by  $\{m+1, \dots, m+p\}$  the rest. With these notations, we see that  $f(x)$  is given by (1.3.1) whenever  $x$  satisfies the set of *constraints*

$$\langle s_j, x \rangle \leq b_j \quad \text{for } j = m+1, \dots, m+p;$$

otherwise,  $f(x) = +\infty$ . Of course, these constraints (usually termed linear, but affine is more correct) define a closed convex polyhedron.

In a word, a polyhedral function is a function which is piecewise affine on its domain, the latter being a closed convex polyhedron. Said otherwise, it is a closed convex function of the form  $\check{f} + I_P$ , where  $\check{f}$  is piecewise affine and  $P$  is a closed convex polyhedron.

**(c) Norms and Distances** It is a direct consequence of the axioms that a norm is a convex function, finite on the whole space (use Definition 1.1.1). More generally, let  $C$  be a nonempty convex set in  $\mathbb{R}^n$  and, with an arbitrary norm  $\|\cdot\|$ , define the *distance function*

$$d_C(x) := \inf \{\|y - x\| : y \in C\}.$$

To establish its convexity, Definition 1.1.1 is again convenient. Take  $\{y_k\}$  and  $\{y'_k\}$  such that, for  $k \rightarrow +\infty$ ,  $\|y_k - x\|$  and  $\|y'_k - x'\|$  tend to  $d_C(x)$  and  $d_C(x')$  respectively. Then form the sequence  $z_k := \alpha y_k + (1 - \alpha) y'_k \in C$  with  $\alpha \in ]0, 1[$ ; pass to the limit for  $k \rightarrow +\infty$  in

$$d_C(\alpha x + (1 - \alpha)x') \leq \|z_k - \alpha x - (1 - \alpha)x'\| \leq \alpha \|y_k - x\| + (1 - \alpha) \|y'_k - x'\|.$$

Here again  $\text{dom } d_C = \mathbb{R}^n$ ; the (lower semi-)continuity of  $d_C$  follows.

Clearly enough,  $d_C = d_{\text{cl } C}$  so, with the help of Proposition III.2.1.8, we see that  $C$ ,  $\text{cl } C$  and  $\text{ri } C$  have the same distance function (associated with the same norm  $\|\cdot\|$ ). In particular,  $d_C$  is 0 on the whole of  $\text{cl } C$ ; the following variant is slightly more accurate, in that it distinguishes between  $\text{int } C$  and  $\text{bd } C$ :

$$D_C(x) := \begin{cases} d_C(x) & \text{if } x \in C^c, \\ -d_{C^c}(x) & \text{if } x \in C, \end{cases}$$

where  $C^c$  is the complement of  $C$  in  $\mathbb{R}^n$ . Assuming that  $C$  and  $C^c$  are both nonempty, it is not particularly difficult to prove that  $D_C$  is convex, finite everywhere, and that

$$\begin{aligned} \text{int } C &= \{x \in \mathbb{R}^n : D_C(x) < 0\}, \\ \text{bd } C &= \{x \in \mathbb{R}^n : D_C(x) = 0\}, \\ (\text{cl } C)^c &= \{x \in \mathbb{R}^n : D_C(x) > 0\}. \end{aligned}$$

**(d) Quadratic Forms** Let  $A : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be a symmetric linear operator. Then the quadratic form

$$f(x) := \frac{1}{2}\langle Ax, x \rangle$$

is a convex function – with  $\text{dom } f = \mathbb{R}^n$  – if and only if  $A$  is positive semi-definite, i.e. its eigenvalues are all nonnegative. Call  $\lambda_1 \geq \dots \geq \lambda_n \geq 0$  these eigenvalues; it is well-known that a basis can be formed with the corresponding eigenvectors, and that as a result,

$$\lambda_n \|x\|^2 \leq \langle Ax, x \rangle \leq \lambda_1 \|x\|^2 \quad \text{for all } x \in \mathbb{R}^n.$$

From the first inequality, direct but somewhat tedious calculations yield, with the notation of (1.1.2):

$$f(\alpha x + (1 - \alpha)x') \leq \alpha f(x) + (1 - \alpha)f(x') - \frac{1}{2}\lambda_n\alpha(1 - \alpha)\|x - x'\|^2.$$

Thus, if  $A$  is positive definite,  $f$  is strongly convex with modulus  $\lambda_n > 0$  (while  $f$  is not even strictly convex when  $A$  is degenerate). A straightforward proof comes also from a general characterization of differentiable strongly convex functions, to be seen below in Theorem 4.1.4 or 4.3.1.

For  $r \geq 0$ , the sublevel-sets of  $f$ :

$$S_r(f) := \{x \in \mathbb{R}^n : \frac{1}{2}\langle Ax, x \rangle \leq r\}$$

are concentric ellipsoids:  $S_{kr}(f) = \sqrt{k}S_r(f)$ . Their common “shape” is given by the eigenvalues of  $A$ . These ellipsoids may be degenerate, in that they contain the subspace  $\text{Ker } A$  (one should rather speak of elliptic cylinders if  $\text{Ker } A \neq \{0\}$ ). However,  $S_r(f)$  is a neighborhood of the origin for  $r > 0$ :

$$S_r(f) \supset B(0, \varepsilon) \quad \text{whenever } \frac{1}{2}\lambda_1\varepsilon^2 \leq r.$$

**(e) Sum of Largest Eigenvalues of a Matrix** Instead of our working space  $\mathbb{R}^n$ , consider the vector space  $S_n(\mathbb{R})$  of symmetric  $n \times n$  matrices. Denote the eigenvalues of  $A \in S_n(\mathbb{R})$  by  $\lambda_1(A) \geq \dots \geq \lambda_n(A)$ , and consider the sum  $f_m$  of the  $m$  largest such eigenvalues ( $m \leq n$  given):

$$S_n(\mathbb{R}) \ni A \mapsto f_m(A) := \sum_{j=1}^m \lambda_j(A).$$

This is a function of  $A$ , finite everywhere. Equip  $S_n(\mathbb{R})$  with the standard dot-product of  $\mathbb{R}^{n \times n}$ :

$$\langle\langle A, B \rangle\rangle := \text{tr } AB = \sum_{i,j=1}^n A_{ij} B_{ij}.$$

The function  $f_m$  turns out to have the following representation:

$$f_m(A) = \sup \{\langle\langle Q Q^\top, A \rangle\rangle : Q \in \Omega\},$$

where  $\Omega := \{Q : Q^\top Q = I_m\}$  is the set of matrices made up of  $m$  orthonormal  $n$ -columns. Indeed,  $\Omega$  is compact and the above supremum is attained at  $Q$  formed with the (normalized) eigenvectors associated with  $\lambda_1, \dots, \lambda_m$ . Keeping Proposition 1.2.8 in mind, this explains that  $f_m$  is convex, as being a supremum of linear functions on  $S_n(\mathbb{R})$ .

Naturally,  $f_1(A)$  is the largest eigenvalue of  $A$ , while  $f_n(A)$  is the trace of  $A$ , a linear function of  $A$ . It follows by taking differences that  $f_n - f_m$  (for example the smallest eigenvalue  $\lambda_n = f_n - f_{n-1}$ ) is a concave function on  $S_n(\mathbb{R})$ .

**(f) Volume of Ellipsoids** Still in the space of symmetric matrices  $S_n(\mathbb{R})$ , define the function

$$A \mapsto f(A) := \begin{cases} \log(\det A^{-1}) & \text{if } A \text{ is positive definite,} \\ +\infty & \text{if not.} \end{cases}$$

It will be seen in §3.1 that the concave finite-valued function  $\lambda_n(\cdot)$  is continuous. The domain of  $f$ , which is the set of  $A \in S_n(\mathbb{R})$  such that  $\lambda_n(A) > 0$ , is therefore open, and even an open convex cone. It turns out that  $f$  is convex. To see it, start from the inequality

$$\det[\alpha A + (1 - \alpha)A'] \geq (\det A)^\alpha (\det A')^{1-\alpha},$$

valid for all symmetric positive definite matrices  $A$  and  $A'$  (and  $\alpha \in ]0, 1[$ ); take the inverse of each side; remember that the inverse of the determinant is the determinant of the inverse; finally, pass to the logarithms.

Geometrically, consider again an ellipsoid

$$E_A := \{x \in \mathbb{R}^n : x^\top A x \leq 1\}$$

where  $A$  is a symmetric positive definite matrix. Up to a positive multiplicative constant (which is the volume of the unit ball  $E_{I_n}$ ), the volume of  $E_A$  is precisely  $\sqrt{\det A^{-1}}$ .

Because  $\text{dom } f$  is open,  $\text{ri dom } f = \text{int dom } f = \text{dom } f$ , which establishes the lower semi-continuity of  $f$  on its domain. Furthermore, suppose  $A_k \rightarrow A$  with  $A$  not positive definite; by continuity of the concave function  $\lambda_n(\cdot)$ ,  $A$  is positive semi-definite and the smallest eigenvalue of  $A_k$  tends to 0:  $f(A_k) \rightarrow +\infty$ . The function  $f$  is closed.

**(g) Epigraphical Hull and Lower-Bound Function of a Convex Set** Given a non-empty convex set  $C \subset \mathbb{R}^n \times \mathbb{R}$ , an interesting question is: when is  $C$  the epigraph of some function  $f \in \text{Conv } \mathbb{R}^n$ ? Let us forget for the moment the convexity issue, which is not really relevant. First, the condition  $f(x) > -\infty$  for all  $x$  means that  $C$  contains no vertical downward half-line:

$$\{r \in \mathbb{R} : (x, r) \in C\} \text{ is minorized for all } x \in \mathbb{R}^n. \quad (1.3.2)$$

A second condition is also obvious:  $C$  must be unbounded from above, more precisely

$$(x, r) \in C \implies (x, r') \in C \text{ for all } r' > r. \quad (1.3.3)$$

The story does not end here, though:  $C$  must have a “closed bottom”, i.e.

$$[(x, r') \in C \text{ and } r' \downarrow r] \implies (x, r) \in C. \quad (1.3.4)$$

This time, we are done: a nonempty set  $C$  satisfying (1.3.2) – (1.3.4) is indeed an epigraph (of a convex function if  $C$  is convex). Alternatively, if  $C$  satisfying (1.3.2), (1.3.3) has its bottom open, i.e.

$$(x, r) \in C \implies (x, r - \varepsilon) \in C \text{ for some } \varepsilon = \varepsilon(x, r) > 0,$$

then  $C$  is a strict epigraph. To cut a long story short: a [strict] epigraph is a union of closed [open] upward half-lines – knowing that we always rule out the value  $-\infty$ .

The next interesting point is to make an epigraph with a given set: the *epigraphical hull* of  $C \subset \mathbb{R}^n \times \mathbb{R}$  is the smallest epigraph containing  $C$ . Its construction involves only rather trivial operations in the ordered set  $\mathbb{R}$ :

- (i) force (1.3.3) by stuffing in everything above  $C$ : for each  $(x, r) \in C$ , add to  $C$  all  $(x, r')$  with  $r' > r$ ;
- (ii) force (1.3.4) by closing the bottom of  $C$ : put  $(x, r)$  in  $C$  whenever  $(x, r') \in C$  with  $r' \rightarrow r$ .

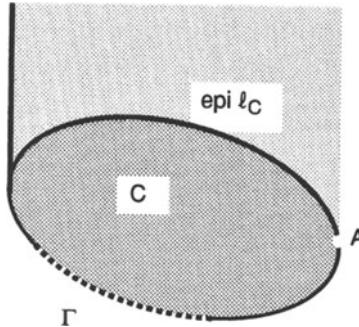
These operations (i), (ii) amount to constructing a function:

$$x \mapsto \ell_C(x) := \inf \{r \in \mathbb{R} : (x, r) \in C\}, \quad (1.3.5)$$

the *lower-bound function* of  $C$ ; clearly enough,  $\text{epi } \ell_C$  is the epigraphical hull of  $C$ . We have that  $\ell_C(x) > -\infty$  for all  $x$  if (and only if)  $C$  satisfies (1.3.2).

The construction of an epigraphical hull is illustrated on Fig. 1.3.1, in which the point  $A$  and the curve  $\Gamma$  are not in  $C$ ; nevertheless, there holds ( $\text{epi}_s$  is the strict epigraph)

$$\text{epi}_s \ell_C \subset C + \{0\} \times \mathbb{R}^+ \subset \text{epi } \ell_C \subset \text{cl}(C + \{0\} \times \mathbb{R}^+). \quad (1.3.6)$$



**Fig. 1.3.1.** The lower-bound function

**Theorem 1.3.1** Let  $C$  be a nonempty subset of  $\mathbb{R}^n \times \mathbb{R}$  satisfying (1.3.2), and let its lower-bound function  $\ell_C$  be defined by (1.3.5).

- (i) If  $C$  is convex, then  $\ell_C \in \text{Conv } \mathbb{R}^n$ ;
- (ii) If  $C$  is closed convex, then  $\ell_C \in \overline{\text{Conv}} \mathbb{R}^n$ .

PROOF. We use the analytical definition (1.1.1). Take arbitrary  $\varepsilon > 0$ ,  $\alpha \in ]0, 1[$  and  $(x_i, r_i) \in C$ ,  $i = 1, 2$  such that

$$r_i \leq \ell_C(x_i) + \varepsilon \quad \text{for } i = 1, 2.$$

When  $C$  is convex,  $(\alpha x_1 + (1 - \alpha)x_2, \alpha r_1 + (1 - \alpha)r_2) \in C$ , hence

$$\ell_C(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha r_1 + (1 - \alpha)r_2 \leq \alpha \ell_C(x_1) + (1 - \alpha)\ell_C(x_2) + \varepsilon.$$

The convexity of  $\ell_C$  follows, since  $\varepsilon > 0$  was arbitrary; (i) is proved.

Now take a sequence  $\{(x_k, \rho_k)\} \subset \text{epi } \ell_C$  converging to  $(x, \rho)$ ; we have to prove  $\ell_C(x) \leq \rho$  (cf. Proposition 1.2.2). By definition of  $\ell_C(x_k)$ , we can select, for each positive integer  $k$ , a real number  $r_k$  such that  $(x_k, r_k) \in C$  and

$$\ell_C(x_k) \leq r_k \leq \ell_C(x_k) + \frac{1}{k} \leq \rho_k + \frac{1}{k}. \quad (1.3.7)$$

We deduce first that  $\{r_k\}$  is bounded from above. Also, when  $\ell_C$  is convex, Proposition 1.2.1 implies the existence of an affine function minorizing  $\ell_C$ :  $\{r_k\}$  is bounded from below.

Extracting a subsequence if necessary, we may assume  $r_k \rightarrow r$ . When  $C$  is closed,  $(x, r) \in C$ , hence  $\ell_C(x) \leq r$ ; but pass to the limit in (1.3.7) to see that  $r \leq \rho$ ; the proof is complete.  $\square$

## 2 Functional Operations Preserving Convexity

It is natural to build up new convex functions from simpler ones, via operations preserving convexity, or even yielding it. This approach goes together with that of

§III.1.2: convex epigraphs can be made up from simpler epigraphs. Here again, proving convexity of the new function will rely either on the analytical definition or on the geometric one, whichever is simpler.

## 2.1 Operations Preserving Closedness

### (a) Positive Combinations of Functions

**Proposition 2.1.1** *Let  $f_1, \dots, f_m$  be in  $\text{Conv } \mathbb{R}^n$  [resp. in  $\overline{\text{Conv}} \mathbb{R}^n$ ],  $t_1, \dots, t_m$  be positive numbers, and assume that there is a point where all the  $f_j$ 's are finite. Then the function*

$$f := \sum_{j=1}^m t_j f_j$$

*is in  $\text{Conv } \mathbb{R}^n$  [resp. in  $\overline{\text{Conv}} \mathbb{R}^n$ ].*

PROOF. The convexity of  $f$  is readily proved from the relation of definition (1.1.1). As for its closedness, start from

$$\liminf_{y \rightarrow x} t_j f_j(y) = t_j \liminf_{y \rightarrow x} f_j(y) \geq t_j f_j(x)$$

(valid for  $t_j > 0$  and  $f_j$  closed); then note that the  $\liminf$  of a sum is not smaller than the sum of  $\liminf$ 's.  $\square$

As an example, let  $f \in \overline{\text{Conv}} \mathbb{R}^n$  and  $C \subset \mathbb{R}^n$  be closed convex, with  $\text{dom } f \cap C \neq \emptyset$ . Then the function  $f + I_C$  of Example 1.3(a) is in  $\overline{\text{Conv}} \mathbb{R}^n$ . This trick can be used to simplify the notation for constrained minimization problems:

$$\inf \{f(x) : x \in C\} \quad \text{and} \quad \inf \{(f + I_C)(x) : x \in \mathbb{R}^n\}$$

are clearly equivalent in the sense that they have the same infimal value and the same solution-set.

### (b) Supremum of Convex Functions

**Proposition 2.1.2** *Let  $\{f_j\}_{j \in J}$  be an arbitrary family of convex [resp. closed convex] functions. If there exists  $x_0$  such that  $\sup_J f_j(x_0) < +\infty$ , then their pointwise supremum*

$$f := \sup \{f_j : j \in J\}$$

*is in  $\text{Conv } \mathbb{R}^n$  [resp. in  $\overline{\text{Conv}} \mathbb{R}^n$ ].*

PROOF. The key property is that a supremum of functions corresponds to an intersection of epigraphs:  $\text{epi } f = \bigcap_{j \in J} \text{epi } f_j$ , which conserves convexity and closedness. The only needed restriction is nonemptiness of this intersection.  $\square$

In a way, this result was already announced by Proposition 1.2.8. It has also been used again and again in the examples of §1.3.

**Example 2.1.3 (Conjugate Function)** Let  $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  be a function not identically  $+\infty$ , minorized by an affine function (i.e., for some  $(s_0, b) \in \mathbb{R}^n \times \mathbb{R}$ ,  $f \geq \langle s_0, \cdot \rangle - b$  on  $\mathbb{R}^n$ ). Then,

$$f^* : \mathbb{R}^n \ni s \mapsto \sup \{\langle s, x \rangle - f(x) : x \in \text{dom } f\}$$

is called the *conjugate function* of  $f$ , to be studied thoroughly in Chap. X. Observe that  $f^*(s_0) \leq b$  and  $f^*(s) > -\infty$  for all  $s$  because  $\text{dom } f \neq \emptyset$ . Thus,  $f^* \in \text{Conv} \mathbb{R}^n$ ; this is true without any further assumption on  $f$ , in particular its convexity or closedness are totally irrelevant here.  $\square$

**Example 2.1.4** Let  $S$  be a nonempty set (not necessarily convex) and take

$$\mathbb{R}^n \ni x \mapsto \varphi_S(x) := \frac{1}{2} [\|x\|^2 - d_S^2(x)],$$

where  $d_S$  is the distance function to  $S$ , associated with the Euclidean norm  $\|\cdot\|$ . A surprising fact is that  $\varphi_S$  is always convex. To see it, develop

$$d_S^2(x) = \inf_{c \in S} \|x - c\|^2 = \|x\|^2 - \sup_{c \in S} [2\langle c, x \rangle - \|c\|^2]$$

to obtain

$$\varphi_S(x) = \sup \{\langle c, x \rangle - \frac{1}{2}\|c\|^2 : c \in S\};$$

$\varphi_S$  thus appears as the pointwise supremum of the affine functions  $\langle c, \cdot \rangle - 1/2\|c\|^2$ , and is closed and convex. In view of the previous example, the reader will realize that  $\varphi_S$  is the conjugate of the function  $1/2\|\cdot\|^2 + I_S$ .  $\square$

### (c) Pre-Composition with an Affine Mapping

**Proposition 2.1.5** Let  $f \in \text{Conv} \mathbb{R}^n$  [resp.  $\text{Conv} \mathbb{R}^n$ ] and let  $A$  be an affine mapping from  $\mathbb{R}^m$  to  $\mathbb{R}^n$  such that  $\text{Im } A \cap \text{dom } f \neq \emptyset$ . Then the function

$$f \circ A : \mathbb{R}^m \ni x \mapsto (f \circ A)(x) = f(A(x))$$

is in  $\text{Conv} \mathbb{R}^m$  [resp.  $\text{Conv} \mathbb{R}^m$ ].

PROOF. Clearly  $(f \circ A)(x) > -\infty$  for all  $x$ , and there exists by assumption  $y = A(x) \in \mathbb{R}^n$  such that  $f(y) < +\infty$ . To check convexity, it suffices to plug the relation

$$A(\alpha x + (1 - \alpha)x') = \alpha A(x) + (1 - \alpha)A(x')$$

into the analytical definition (1.1.1) of convexity. As for closedness, it comes readily from the continuity of  $A$  when  $f$  is itself closed.  $\square$

**Example 2.1.6** With  $f$  (closed) convex on  $\mathbb{R}^n$ , take  $x_0 \in \text{dom } f$ ,  $d \in \mathbb{R}^n$  and define

$$A : \mathbb{R} \ni t \mapsto A(t) = x_0 + td;$$

this  $A$  is affine, its linear part is  $t \mapsto A_0t := td$ . The resulting  $f \circ A$  appears as (a parametrization of) the restriction of  $f$  along the line  $x_0 + \mathbb{R}d$ , which meets  $\text{dom } f$  (at  $x_0$ ).

This operation is often used in applications: think for example of the line-search problem, considered in §II.3. Even from a theoretical point of view, the one-dimensional traces of  $f$  are important, in that  $f$  itself inherits many of their properties; Proposition 1.2.5 gives an instance of this phenomenon.  $\square$

**Remark 2.1.7** With relation to this operation on  $f \in \text{Conv } \mathbb{R}^n$  [resp.  $\overline{\text{Conv}} \mathbb{R}^n$ ], call  $V$  the subspace parallel to  $\text{aff dom } f$ . Then, fix  $x_0 \in \text{dom } f$  and define the convex function  $f_0 \in \text{Conv } V$  [resp.  $\overline{\text{Conv}} V$ ] by

$$f_0(y) := f(x_0 + y) \quad \text{for all } y \in V.$$

This new function is obtained from  $f$  by a simple translation, *composed with a restriction* (from  $\mathbb{R}^n$  to  $V$ ). As a result,  $\text{dom } f_0$  is now full-dimensional (in  $V$ ), the relative topology relevant for  $f_0$  is the standard topology of  $V$ . This trick is often useful and explains why “flat” domains, instead of full-dimensional, create little difficulties.  $\square$

#### (d) Post-Composition with an Increasing Convex Function

**Proposition 2.1.8** Let  $f \in \text{Conv } \mathbb{R}^n$  [resp.  $\overline{\text{Conv}} \mathbb{R}^n$ ] and let  $g \in \text{Conv } \mathbb{R}$  [resp.  $\overline{\text{Conv}} \mathbb{R}$ ] be increasing. Assume that there is  $x_0 \in \mathbb{R}^n$  such that  $f(x_0) \in \text{dom } g$ , and set  $g(+\infty) := +\infty$ . Then the composite function  $g \circ f : x \mapsto g(f(x))$  is in  $\text{Conv } \mathbb{R}^n$  [resp. in  $\overline{\text{Conv}} \mathbb{R}^n$ ].

PROOF. It suffices to check the inequalities of definition: (1.1.1) for convexity, (1.2.3) for closedness.  $\square$

The exponential  $g(t) := \exp t$  is convex increasing, its domain is the whole line, so  $\exp f(x)$  is a [closed] convex function of  $x \in \mathbb{R}^n$  whenever  $f$  is [closed] convex. A function  $f : \mathbb{R}^n \rightarrow ]0, +\infty]$  is called *logarithmically convex* when  $\log f \in \text{Conv } \mathbb{R}^n$  (we set again  $\log(+\infty) = +\infty$ ). Because  $f = \exp \log f$ , a logarithmically convex function is convex.

As another application, the square of an arbitrary nonnegative convex function (for example a norm) is convex: post-compose it by the function  $g(t) = (\max\{0, t\})^2$ .

## 2.2 Dilations and Perspectives of a Function

For a convex function  $f$  and  $u > 0$ , the function

$$f_u : \mathbb{R}^n \ni x \mapsto f_u(x) = uf(x/u)$$

is again convex. This comes from Propositions 2.1.1 and 2.1.5 but can also be seen geometrically: since  $f_u(x)/u = f(x/u)$ , the epigraphs and sublevel-sets are related by

$$\text{epi } f_u = u \text{ epi } f, \quad \text{epi}_s f_u = u \text{ epi}_s f, \quad S_r(f_u) = u S_{r/u}(f),$$

which express that  $f_u$  is a “dilated version” of  $f$ .

More interesting, however, is to study  $f_u$  as a function of *both* variables  $x$  and  $u$ , i.e. to consider the set of all dilations of  $f$ . We therefore define the *perspective* of  $f$  as the function from  $\mathbb{R} \times \mathbb{R}^n$  to  $\mathbb{R} \cup \{+\infty\}$  given by

$$\tilde{f}(u, x) := \begin{cases} uf(x/u) & \text{if } u > 0, \\ +\infty & \text{if not.} \end{cases}$$

**Proposition 2.2.1** If  $f \in \text{Conv } \mathbb{R}^n$ , its perspective  $\tilde{f}$  is in  $\text{Conv } \mathbb{R}^{n+1}$ .

PROOF. Here also, it is better to look at  $\tilde{f}$  with “geometric glasses”:

$$\begin{aligned}\text{epi } \tilde{f} &= \{(u, x, r) \in \mathbb{R}_*^+ \times \mathbb{R}^n \times \mathbb{R} : f(x/u) \leq r/u\} \\ &= \{u(1, x', r') : u > 0, (x', r') \in \text{epi } f\} \\ &= \cup_{u>0} \{u(\{1\} \times \text{epi } f)\} = \mathbb{R}_*^+(\{1\} \times \text{epi } f)\end{aligned}$$

and  $\text{epi } \tilde{f}$  is therefore a convex cone.  $\square$

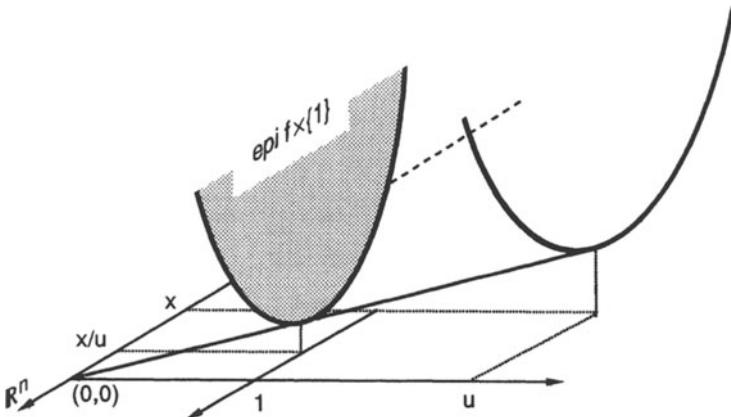


Fig. 2.2.1. The perspective of a convex function

Figure 2.2.1 illustrates the construction of  $\text{epi } \tilde{f}$ , as given in the above proof. Embed  $\text{epi } f$  into  $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}$ , where the first  $\mathbb{R}$  represents the extra variable  $u$ ; shift it horizontally by one unit; finally, take the positive multiples of the result. Observe that, following the same technique, we obtain

$$\text{dom } \tilde{f} = \mathbb{R}_*^+(\{1\} \times \text{dom } f). \quad (2.2.1)$$

Another observation is that, by construction,  $\text{epi } \tilde{f}$  [resp.  $\text{dom } \tilde{f}$ ] does not contain the origin of  $\mathbb{R} \times \mathbb{R}^n \times \mathbb{R}$  [resp.  $\mathbb{R} \times \mathbb{R}^n$ ].

Convexity of a perspective-function is an important property, which we will use later in the following way. For fixed  $x_0 \in \text{dom } f$ , the function  $d \mapsto f(x_0 + d) - f(x_0)$  is obviously convex, so its perspective

$$r(u, d) := u[f(x_0 + d/u) - f(x_0)] \quad (\text{for } u > 0) \quad (2.2.2)$$

is also convex with respect to the couple  $(u, d) \in \mathbb{R}_*^+ \times \mathbb{R}^n$ . Up to the simple change of variable  $u \mapsto t = 1/u$ , we recognize a difference quotient.

The next natural question is the closedness of a perspective-function: admitting that  $f$  itself is closed, troubles can still be expected at  $u = 0$ , where we have brutally set  $f(0, \cdot) = +\infty$  (possibly not the best idea . . .) A relatively simple calculation of  $\text{cl } \tilde{f}$  is in fact given by Proposition 1.2.5:

**Proposition 2.2.2** *Let  $f \in \text{Conv} \mathbb{R}^n$  and let  $x' \in \text{ri dom } f$ . Then the closure  $\text{cl } \tilde{f}$  of its perspective is given as follows:*

$$(\text{cl } \tilde{f})(u, x) = \begin{cases} uf(x/u) & \text{if } u > 0, \\ \lim_{\alpha \downarrow 0} \alpha f(x' - x + x/\alpha) & \text{if } u = 0, \\ +\infty & \text{if } u < 0. \end{cases}$$

PROOF. Suppose first  $u < 0$ . For any  $x$ , it is clear that  $(u, x)$  is outside  $\text{cl dom } \tilde{f}$  and, in view of (1.2.8),  $\text{cl } \tilde{f}(u, x) = +\infty$ .

Now let  $u \geq 0$ . Using (2.2.1), the assumption on  $x'$  and the results of §III.2.1, we see that  $(1, x') \in \text{ri dom } \tilde{f}$ , so Proposition 1.2.5 allows us to write

$$\begin{aligned} (\text{cl } \tilde{f})(u, x) &= \lim_{\alpha \downarrow 0} \tilde{f}((u, x) + \alpha[(1, x') - (u, x)]) \\ &= \lim_{\alpha \downarrow 0} [u + \alpha(1-u)] f\left(\frac{x+\alpha(x'-x)}{u+\alpha(1-u)}\right). \end{aligned}$$

If  $u = 1$ , this reads  $\text{cl } \tilde{f}(1, x) = \text{cl } f(x) = f(x)$  (because  $f$  is closed); if  $u = 0$ , we just obtain our claimed relation.  $\square$

**Remark 2.2.3** Observe that the behaviour of  $\tilde{f}(u, \cdot)$  for  $u \downarrow 0$  just depends on the behaviour of  $f$  at infinity. If  $x = 0$ , we have

$$\text{cl } \tilde{f}(0, 0) = \lim_{\alpha \downarrow 0} \alpha f(x') = 0 \quad [f(x') < +\infty].$$

For  $x \neq 0$ , suppose for example that  $\text{dom } f$  is bounded; then  $f(x' - x + x/\alpha) = +\infty$  for  $\alpha$  small enough and  $\text{cl } \tilde{f}(0, x) = +\infty$ . On the other hand, when  $\text{dom } f$  is unbounded,  $\text{cl } \tilde{f}(0, \cdot)$  may assume finite values if, at infinity,  $f$  does not increase too fast.

For another illustration, we apply here Proposition 2.2.2 to the perspective-function  $r$  of (2.2.2). Assuming  $x_0 \in \text{ri dom } f$ , we can take  $d' = 0$  – which is in the relative interior of the function  $d \mapsto f(x_0 + d) - f(x_0)$  – to obtain

$$(\text{cl } r)(0, d) = \lim_{\tau \rightarrow +\infty} \frac{f(x_0 - d + \tau d) - f(x_0)}{\tau}.$$

Because  $(\tau - 1)/\tau \rightarrow 1$  for  $\tau \rightarrow +\infty$ , the last limit can also be written (in  $\mathbb{R} \cup \{+\infty\}$ )

$$(\text{cl } r)(0, d) = \lim_{t \rightarrow +\infty} \frac{f(x_0 + td) - f(x_0)}{t}.$$

We will return to all this in §3.2 below.  $\square$

## 2.3 Infimal Convolution

Starting from two functions  $f_1$  and  $f_2$ , form the set  $\text{epi } f_1 + \text{epi } f_2 \subset \mathbb{R}^n \times \mathbb{R}$ :

$$C := \{(x_1 + x_2, r_1 + r_2) : r_j \geq f_j(x_j) \text{ for } j = 1, 2\}.$$

Under a suitable minorization property, this  $C$  has a lower-bound function  $\ell_C$  as in (1.3.5):

$$\ell_C(x) = \inf \{r_1 + r_2 : r_j \geq f_j(x_j) \text{ for } j = 1, 2, x_1 + x_2 = x\}.$$

In the above minimization problem, the variables are  $r_1, r_2, x_1, x_2$ , but the  $r_j$ 's can be eliminated; in fact,  $\ell_C$  can be defined as follows.

**Definition 2.3.1** Let  $f_1$  and  $f_2$  be two functions from  $\mathbb{R}^n$  to  $\mathbb{R} \cup \{+\infty\}$ . Their *infimal convolution* is the function from  $\mathbb{R}^n$  to  $\mathbb{R} \cup \{\pm\infty\}$  defined by

$$(f_1 \downarrow f_2)(x) := \inf \{f_1(x_1) + f_2(x_2) : x_1 + x_2 = x\} = \inf_{y \in \mathbb{R}^n} [f_1(y) + f_2(x - y)]. \quad (2.3.1)$$

We will also call “infimal convolution” the *operation* expressed by (2.3.1). It is called *exact* at  $x = \bar{x}_1 + \bar{x}_2$  when the infimum is attained at  $(\bar{x}_1, \bar{x}_2)$ , not necessarily unique.  $\square$

We refer to Remark I.2.1.4 for an explanation and some comments on the terminology “infimal convolution”. To exclude the undesired value  $-\infty$  from the range of an inf-convolution, an additional assumption is obviously needed: in one dimension, the infimal convolution of the functions  $x$  and  $-x$  is identically  $-\infty$ . Our next result proposes a convenient such assumption.

**Proposition 2.3.2** Let the functions  $f_1$  and  $f_2$  be in  $\text{Conv } \mathbb{R}^n$ . Suppose that they have a common affine minorant: for some  $(s, b) \in \mathbb{R}^n \times \mathbb{R}$ ,

$$f_j(x) \geq \langle s, x \rangle - b \quad \text{for } j = 1, 2 \text{ and all } x \in \mathbb{R}^n.$$

Then their infimal convolution is also in  $\text{Conv } \mathbb{R}^n$ .

PROOF. For arbitrary  $x \in \mathbb{R}^n$  and  $x_1, x_2$  such that  $x_1 + x_2 = x$ , we have by assumption

$$f_1(x_1) + f_2(x_2) \geq \langle s, x \rangle - 2b > -\infty,$$

and this inequality extends to the infimal value  $(f_1 \downarrow f_2)(x)$ .

On the other hand, it suffices to choose particular values  $x_j \in \text{dom } f_j$ ,  $j = 1, 2$ , to obtain the point  $x_1 + x_2 \in \text{dom}(f_1 \downarrow f_2)$ . Finally, the convexity of  $f_1 \downarrow f_2$  results from the convexity of a lower-bound function, as seen in §1.3(g).  $\square$

**Remark 2.3.3** To prove that an inf-convolution of convex functions is convex, one can also show the following relation between strict epigraphs:

$$\text{epi}_s(f_1 \downarrow f_2) = \text{epi}_s f_1 + \text{epi}_s f_2. \quad (2.3.2)$$

In fact,  $(x, r) \in \text{epi}_s(f_1 \downarrow f_2)$  if and only if there is  $\varepsilon > 0$  such that

$$f_1(x_1) + f_2(x_2) = r + \varepsilon \quad \text{for some } x_1 \text{ and } x_2 \text{ adding up to } x.$$

This is equivalent to

$$f_j(x_j) < r_j \quad \text{for some } (x_1, r_1) \text{ and } (x_2, r_2) \text{ adding up to } (x, r)$$

(set  $r_j := f_j(x_j) + \varepsilon/2$  for  $j = 1, 2$ , to show the “ $\Rightarrow$ ” direction). This last property holds if and only if  $(x, r) \in \text{epi}_s f_1 + \text{epi}_s f_2$ .

This proof explains why the infimal convolution is sometimes called the (strict) *epigraphic addition*.  $\square$

Similarly to (2.3.2), we have by construction

$$\text{dom}(f_1 \downarrow f_2) = \text{dom } f_1 + \text{dom } f_2.$$

Let us mention some immediate properties of the infimal convolution:

$$f_1 \downarrow f_2 = f_2 \downarrow f_1 \quad (\text{commutativity}) \tag{2.3.3}$$

$$(f_1 \downarrow f_2) \downarrow f_3 = f_1 \downarrow (f_2 \downarrow f_3) \quad (\text{associativity}) \tag{2.3.4}$$

$$f \downarrow I_{\{0\}} = f \quad (\text{existence of a neutral element in } \text{Conv } \mathbb{R}^n) \tag{2.3.5}$$

$$f_1 \leqslant f_2 \implies f_1 \downarrow g \leqslant f_2 \downarrow g \quad (\downarrow \text{preserves the order}).$$

With relation to (2.3.3), (2.3.4), more than two functions can of course be inf-convolved:

$$(f_1 \downarrow \cdots \downarrow f_m)(x) = \inf \left\{ \sum_{j=1}^m f_j(x_j) : \sum_{j=1}^m x_j = x \right\}.$$

**Remark 2.3.4** If  $C_1$  and  $C_2$  are nonempty convex sets in  $\mathbb{R}^n$ , then

$$I_{C_1} \downarrow I_{C_2} = I_{C_1 + C_2}.$$

This is due to the additional nature of the inf-convolution, and can also be checked directly; but it leads us to an important observation: since the sum of two closed sets may not be closed, an infimal convolution *need not be closed*, even if it is constructed from two closed functions and if it is exact everywhere.  $\square$

**Example 2.3.5** Let  $C$  be a nonempty convex subset of  $\mathbb{R}^n$  and  $\|\cdot\|$  an arbitrary norm. Then

$$I_C \downarrow \|\cdot\| = d_C,$$

which confirms the convexity of the distance function  $d_C$ . It also shows that inf-convolving two non-closed functions ( $C$  need not be closed) may result in a closed function.  $\square$

**Example 2.3.6** Let  $f$  be an arbitrary convex function minorized by some affine function with slope  $s$ . Taking an affine function  $g = \langle s, \cdot \rangle - b$ , we obtain

$$f \downarrow g = g - c$$

where  $c$  is a constant:  $c = \sup_y [\langle s, y \rangle - f(y)]$ . Note: we have already encountered in Example 2.1.3  $c = f^*(s)$ , the value at  $s$  of the conjugate of  $f$ .

Take in particular a constant function for  $g$ : assuming  $f$  bounded below,

$$-g = \bar{f} := \inf_y f(y).$$

Then

$$f \downarrow (-\bar{f}) = 0.$$

Do not believe, however, that the infimal convolution provides  $\text{Conv } \mathbb{R}^n$  with the structure of a commutative group: in view of (2.3.5), the 0-function is not the neutral element!  $\square$

**Example 2.3.7** We have seen (Proposition 1.2.1) that a convex function is indeed minorized by some affine function. The dilated versions  $f_u = uf(\cdot/u)$  of a given convex function  $f$  are minorized by some affine function with a slope *independent* of  $u > 0$ , and can be inf-convolved by each other. We obtain

$$f_u \downarrow f_{u'} = f_{u+u'};$$

the quickest way to prove this formula is probably to use (2.3.2), knowing that  $\text{epi}_s f_u = u \text{ epi}_s f$ . In particular, inf-convolving  $m$  times a function with itself gives a sort of mean-value formula:

$$\frac{1}{m}(f \downarrow \cdots \downarrow f)(x) = f\left(\frac{1}{m}x\right).$$

Observe how a perspective-function gives a meaning to a non-integer number of self-inf-convolutions.  $\square$

**Example 2.3.8** Consider two quadratic forms

$$f_j(x) = \frac{1}{2}\langle A_j x, x \rangle \quad \text{for } j = 1, 2,$$

with  $A_1$  and  $A_2$  symmetric positive definite. Expressing their infimal convolution as

$$\frac{1}{2} \inf_y [\langle A_1 y, y \rangle + \langle A_2(x - y), x - y \rangle],$$

the minimum can be explicitly worked out, to give  $(f_1 \downarrow f_2)(x) = \frac{1}{2}\langle A_{12}x, x \rangle$ , where

$$A_{12} := (A_1^{-1} + A_2^{-1})^{-1}.$$

This formula has an interesting physical interpretation: consider an electrical circuit made up of two generalized resistors  $A_1$  and  $A_2$  connected in parallel. A given current-vector  $i \in \mathbb{R}^n$  is distributed among the two branches ( $i = i_1 + i_2$ ), in such a way that the dissipated power  $\langle A_1 i_1, i_1 \rangle + \langle A_2 i_2, i_2 \rangle$  is minimal (this is Maxwell's variational principle); see Fig. 2.3.1. In other words, if  $i = \bar{i}_1 + \bar{i}_2$  is the real current distribution, we must have

$$\langle A_1 \bar{i}_1, \bar{i}_1 \rangle + \langle A_2 \bar{i}_2, \bar{i}_2 \rangle = \inf_{i_1+i_2=i} (\langle A_1 i_1, i_1 \rangle + \langle A_2 i_2, i_2 \rangle).$$

The unique distribution  $(\bar{i}_1, \bar{i}_2)$  is thus characterized by the formulae

$$A_1 \bar{i}_1 = A_2 \bar{i}_2 = A_{12} i, \tag{2.3.6}$$

from which it follows that

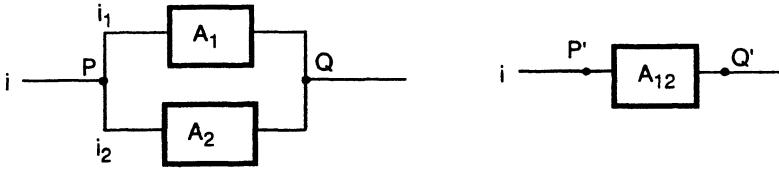


Fig. 2.3.1. Equivalent resistors

$$\langle A_1 \bar{i}_1, \bar{i}_1 \rangle + \langle A_2 \bar{i}_2, \bar{i}_2 \rangle = \langle A_{12} i, i \rangle.$$

Thus, \$A\_{12}\$ plays the role of a generalized resistor equivalent to \$A\_1\$ and \$A\_2\$ connected in parallel; when \$n = 1\$, we get the more familiar relation \$1/r = 1/r\_1 + 1/r\_2\$ between ordinary resistances \$r\_1\$ and \$r\_2\$. Note an interpretation of the optimality (or equilibrium) condition (2.3.6). The voltage between \$P\$ and \$Q\$ [resp. \$P'\$ and \$Q'\$] on Fig. 2.3.1, namely \$A\_1 \bar{i}\_1 = A\_2 \bar{i}\_2\$ [resp. \$A\_{12} i\$], is independent of the path chosen: either through \$A\_1\$, or through \$A\_2\$, or by construction through \$A\_{12}\$.

The above example of two convex quadratic functions can be extended to general functions, and it gives an economic interpretation of the infimal convolution: let \$f\_1(x)\$ [resp. \$f\_2(x)\$] be the cost of producing \$x\$ by some production unit \$U\_1\$ [resp. \$U\_2\$]. If we want to distribute optimally the production of a given \$x\$ between \$U\_1\$ and \$U\_2\$, we have to solve the minimization problem (2.3.1). \$\square\$

**Remark 2.3.9** In Example III.1.2.6, we have seen two kinds of differences between sets, which may be applied to epigraphs. One difference, \$C\_1 - C\_2 = C\_1 + (-C\_2)\$, leads nowhere: the opposite of an epigraph is not an epigraph. On the other hand, it is not too difficult to see that the star-difference of two epigraphs is again an epigraph; it therefore corresponds to an operation with convex functions, namely the *deconvolution*, or epigraphic star-difference:

$$(f_1 \bar{\vee} f_2)(x) := \sup \{f_1(x + y) - f_2(y) : y \in \text{dom } f_2\},$$

Being a supremum of convex functions, the result is a convex function provided that

$$[\text{epi}(f_1 \bar{\vee} f_2) =] \text{ epi } f_1 * \text{epi } f_2 \neq \emptyset.$$

In the language of function-values, this means that, for some \$(x\_0, r\_0) \in \mathbb{R}^n \times \mathbb{R}\$:

$$f_1(x) \leq f_2(x - x_0) + r_0 \quad \text{for all } x \in \mathbb{R}^n.$$

In words: \$f\_1\$ must not be too larger than \$f\_2\$.

Indeed, the above operation can be seen to a great extent as the inverse operation of the inf-convolution. It goes without saying that the deconvolution is not commutative. A detail is worth mentioning, though: in contrast to the inf-convolution, \$f\_1 \bar{\vee} f\_2\$ is now a supremum; by virtue of Proposition 2.1.2, it is therefore closed when \$f\_1\$ is closed. \$\square\$

## 2.4 Image of a Function Under a Linear Mapping

Consider a constrained optimization problem, formally written as

$$\inf_{u \in U} \{\varphi(u) : c(u) \leqslant x\}, \quad (2.4.1)$$

where the optimization variable is  $u$ , the right-hand side  $x$  being considered as a parameter taken in some ordered set  $X$ . The optimal value in such a problem is then a function of  $x$ , characterized by the triple  $(U, \varphi, c)$ , and taking its values in  $\mathbb{R} \cup \{\pm\infty\}$ . In convex analysis and optimization, this is an important function, usually called the *value* function, or *marginal* function, or *perturbation* function, or *primal* function, etc.

Several variants of (2.4.1) are possible: we may encounter equality constraints, some constraints may be included in the objective via an indicator function, etc. A convenient unified formulation is the following:

**Definition 2.4.1** Let  $A : \mathbb{R}^m \rightarrow \mathbb{R}^n$  be linear and let  $g : \mathbb{R}^m \rightarrow \mathbb{R} \cup \{+\infty\}$ . The *image* of  $g$  under  $A$  is the function  $Ag : \mathbb{R}^n \rightarrow \mathbb{R} \cup \pm\infty$  defined by

$$(Ag)(x) := \inf \{g(y) : Ay = x\} \quad (2.4.2)$$

(here as always,  $\inf \emptyset = +\infty$ ). □

The terminology comes from the case of an indicator function: when  $g = I_C$ , with  $C$  nonempty in  $\mathbb{R}^m$ , (2.4.2) writes

$$(Ag)(x) = \begin{cases} 0 & \text{if } x = Ay \text{ for some } y \in C, \\ +\infty & \text{otherwise.} \end{cases}$$

In other words,  $Ag = I_{A(C)}$  is the indicator function of the image of  $C$  under  $A$  (and we know from Proposition III.1.2.4 that this image is convex when  $C$  is convex).

Even if  $U$  and  $X$  in (2.4.1) are Euclidean spaces, we seem to limit the generality when passing to (2.4.2), since only linear constraints are considered. Actually, (2.4.1) can be put in the form (2.4.2): with  $X = \mathbb{R}^n$  and  $y = (u, v) \in U \times X = \mathbb{R}^m$ , define  $Ay := v$  and  $g(y) := \varphi(u) + I_C(y)$ , where

$$C := \{y = (u, v) \in \mathbb{R}^m : c(u) \leqslant v\}. \quad (2.4.3)$$

Note that conversely, (2.4.2) can be put in the form (2.4.1) via an analogous trick turning its equality constraints into inequalities.

**Theorem 2.4.2** Let  $g$  of Definition 2.4.1 be in  $\text{Conv } \mathbb{R}^m$ . Assume also that, for all  $x \in \mathbb{R}^n$ ,  $g$  is bounded from below on the inverse image

$$A^{-1}(x) = \{y \in \mathbb{R}^m : Ay = x\}.$$

Then  $Ag \in \text{Conv } \mathbb{R}^n$ .

PROOF. By assumption,  $Ag$  is nowhere  $-\infty$ ; also,  $(Ag)(x) < +\infty$  whenever  $x = Ay$ , with  $y \in \text{dom } g$ . Now consider the extended operator

$$A' : \mathbb{R}^m \times \mathbb{R} \ni (y, r) \mapsto A'(y, r) := (Ay, r) \in \mathbb{R}^m \times \mathbb{R}.$$

The set  $A'(\text{epi } g) =: C$  is convex in  $\mathbb{R}^n \times \mathbb{R}$ , let us compute its lower-bound function (1.3.5): for given  $x \in \mathbb{R}^n$ ,

$$\begin{aligned}\inf_r \{r : (x, r) \in C\} &= \inf_{y, r} \{r : Ay = x \text{ and } g(y) \leq r\} \\ &= \inf_y \{g(y) : Ay = x\} = (Ag)(x),\end{aligned}$$

and this proves the convexity of  $Ag = \ell_C$ .  $\square$

Usually,  $\overline{A}(x)$  contains several points – it is an affine manifold of  $\mathbb{R}^n$  – and  $Ag(x)$  selects one giving the least value of  $g$  (admitting that (2.4.2) has a solution). If  $A$  is invertible,  $Ag = g \circ A^{-1}$ ; more generally, the above proof discloses the following interpretation:  $\text{epi}(Ag)$  is the epigraphical hull of the inverse image  $(A')(\text{epi } g)$  (a convex set in  $\mathbb{R}^n \times \mathbb{R}$ ).

**Corollary 2.4.3** *Let (2.4.1) have the following form:  $U = \mathbb{R}^p$ ;  $\varphi \in \text{Conv } \mathbb{R}^p$ ;  $X = \mathbb{R}^n$  is equipped with the canonical basis; the mapping  $c$  has its components  $c_j \in \text{Conv } \mathbb{R}^p$  for  $j = 1, \dots, n$ . Suppose also that the optimal value is  $> -\infty$  for all  $x \in \mathbb{R}^n$ , and that*

$$\text{dom } \varphi \cap \text{dom } c_1 \cap \dots \cap \text{dom } c_n \neq \emptyset. \quad (2.4.4)$$

*Then the value function*

$$v_{\varphi, c}(x) := \inf \{\varphi(u) : c_j(u) \leq x_j \text{ for } j = 1, \dots, n\}$$

*is in  $\text{Conv } \mathbb{R}^n$ .*

**PROOF.** Note first that we have assumed  $v_{\varphi, c}(x) > -\infty$  for all  $x$ . Take  $u_0$  in the set (2.4.4) and set  $M := \max_j c_j(u_0)$ ; then take  $x_0 := (M, \dots, M) \in \mathbb{R}^n$ , so that  $v_{\varphi, c}(x_0) \leq \varphi(u_0) < +\infty$ . Knowing that  $v_{\varphi, c}$  is an image-function, we just have to prove the convexity of the set (2.4.3); but this in turn comes immediately from the convexity of each  $c_j$ .  $\square$

Taking the image of a convex function under a linear mapping can be used as a mould to describe a number of other operations – (2.4.1) is indeed one of them. An example is the infimal convolution of §2.3: with  $f_1$  and  $f_2$  in  $\text{Conv } \mathbb{R}^n$ , define  $g \in \text{Conv}(\mathbb{R}^n \times \mathbb{R}^n)$  by

$$g(x_1, x_2) := f_1(x_1) + f_2(x_2)$$

and  $A : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}^n$  by

$$A(x_1, x_2) := x_1 + x_2.$$

Then we have  $Ag = f_1 \downarrow f_2$  and (2.3.1) is put in the form (2.4.2). Incidentally, this shows that an image of a closed function need not be closed.

Another example has lots of practical applications: the *marginal* function of  $g \in \text{Conv}(\mathbb{R}^n \times \mathbb{R}^m)$  is

$$\mathbb{R}^n \ni x \mapsto \gamma(x) := \inf \{g(x, y) : y \in \mathbb{R}^m\}.$$

This is the image of  $g$  under the linear mapping projecting each  $(x, y) \in \mathbb{R}^n \times \mathbb{R}^m$  onto  $x \in \mathbb{R}^n$ . It is therefore convex if  $g$  is bounded below on the set  $\{x\} \times \mathbb{R}^m$  for all  $x \in \mathbb{R}^n$ . Geometrically, a marginal function is given by Fig. 2.4.1, which explains why convexity is preserved: the strict epigraph of  $\gamma$  is the projection onto  $\mathbb{R}^n \times \mathbb{R}$  of the strict epigraph of  $g$  ( $\subset \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}$ ). Therefore,  $\text{epi}_s \gamma$  is also the image of a convex set under a linear mapping; see again Example III.1.2.5.

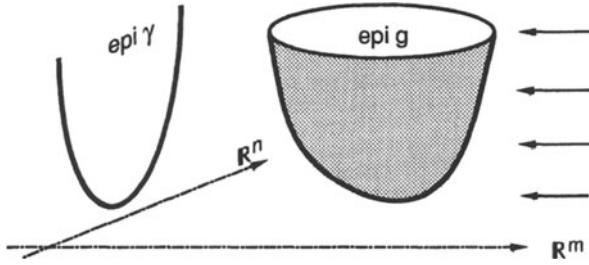


Fig. 2.4.1. The shadow of a convex epigraph

As seen in §2.1(b), supremization preserves convexity. Here, if  $g(\cdot, y)$  were concave for each  $y$ ,  $\gamma$  would therefore be concave: the convexity of  $\gamma$  is a little bit surprising. Needless to say, it is the convexity of  $g$  with respect to the *couple* of variables  $x$  and  $y$  that is crucial.

## 2.5 Convex Hull and Closed Convex Hull of a Function

Given a (nonconvex) function  $g$ , a natural idea coming from §III.1.3 is to take the convex hull  $\text{co epi } g$  of its epigraph. This gives a convex set, which is not an epigraph, but which can be made so by “closing its bottom” via its lower-bound function (1.3.5). As seen in §III.1.3, there are several ways of constructing a convex hull; the next result exploits them, and uses the unit simplex of  $\mathbb{R}^k$ :

$$\Delta_k := \left\{ (\alpha_1, \dots, \alpha_k) \in \mathbb{R}^k : \sum_{j=1}^k \alpha_j = 1, \alpha_j \geq 0 \text{ for } j = 1, \dots, k \right\}. \quad (2.5.1)$$

**Proposition 2.5.1** *Let  $g : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ , not identically  $+\infty$ , be minorized by an affine function: for some  $(s, b) \in \mathbb{R}^n \times \mathbb{R}$ ,*

$$g(x) \geq \langle s, x \rangle - b \quad \text{for all } x \in \mathbb{R}^n. \quad (2.5.2)$$

*Then, the following three functions  $f_1$ ,  $f_2$  and  $f_3$  are convex and coincide on  $\mathbb{R}^n$ :*

$$\begin{aligned} f_1(x) &:= \inf \{r : (x, r) \in \text{co epi } g\}, \\ f_2(x) &:= \sup \{h(x) : h \in \text{Conv } \mathbb{R}^n, h \leq g\}. \\ f_3(x) &:= \inf \left\{ \sum_{j=1}^k \alpha_j g(x_j) : k = 1, 2, \dots \right. \\ &\quad \left. \alpha \in \Delta_k, x_j \in \text{dom } g, \sum_{j=1}^k \alpha_j x_j = x \right\}. \end{aligned} \quad (2.5.3)$$

PROOF. We denote by  $\Gamma$  the family of convex functions minorizing  $g$ . By assumption,  $\Gamma \neq \emptyset$ ; then the convexity of  $f_1$  results from §1.3(g).

[ $f_2 \leq f_1$ ] Consider the epigraph of any  $h \in \Gamma$ : its lower-bound function  $\ell_{\text{epi } h}$  is  $h$  itself; besides, it contains  $\text{epi } g$ , and  $\text{co}(\text{epi } g)$  as well (see Proposition III.1.3.4). In a word, there holds

$$h = \ell_{\text{epi } h} \leq \ell_{\text{co epi } g} = f_1$$

and we conclude  $f_2 \leq f_1$  since  $h$  was arbitrary in  $\Gamma$ .

[ $f_3 \leq f_2$ ] We have to prove  $f_3 \in \Gamma$ , and the result will follow by definition of  $f_2$ ; clearly  $f_3 \leq g$  (take  $\alpha \in \Delta_1$ !), so it suffices to establish  $f_3 \in \text{Conv } \mathbb{R}^n$ . First, with  $(s, b)$  of (2.5.2) and all  $x, \{x_j\}$  and  $\{\alpha_j\}$  as described by (2.5.3),

$$\sum_{j=1}^k \alpha_j g(x_j) \geq \sum_{j=1}^k \alpha_j (\langle s, x_j \rangle - b) = \langle s, x \rangle - b;$$

hence  $f_3$  is minorized by the affine function  $\langle s, \cdot \rangle - b$ . Now, take two points  $(x, r)$  and  $(x', r')$  in the strict epigraph of  $f_3$ . By definition of  $f_3$ , there are  $k, \{\alpha_j\}, \{x_j\}$  as described in (2.5.3), and likewise  $k', \{\alpha'_j\}, \{x'_j\}$ , such that

$$\sum_{j=1}^k \alpha_j g(x_j) < r \quad \text{and likewise} \quad \sum_{j=1}^{k'} \alpha'_j g(x'_j) < r'.$$

For arbitrary  $t \in ]0, 1[$ , we obtain by convex combination

$$\sum_{j=1}^k t \alpha_j g(x_j) + \sum_{j=1}^{k'} (1-t) \alpha'_j g(x'_j) < tr + (1-t)r'.$$

Observe that

$$\sum_{j=1}^k t \alpha_j x_j + \sum_{j=1}^{k'} (1-t) \alpha'_j x'_j = tx + (1-t)x',$$

i.e. we have in the left-hand side a convex decomposition of  $tx + (1-t)x'$  in  $k + k'$  elements; therefore, by definition of  $f_3$ :

$$f_3(tx + (1-t)x') \leq \sum_{j=1}^k t \alpha_j g(x_j) + \sum_{j=1}^{k'} (1-t) \alpha'_j g(x'_j)$$

and we have proved that  $\text{epi}_s f_3$  is a convex set:  $f_3$  is convex.

[ $f_1 \leq f_3$ ] Let  $x \in \mathbb{R}^n$  and take an arbitrary convex decomposition  $x = \sum_{j=1}^k \alpha_j x_j$ , with  $\alpha_j$  and  $x_j$  as described in (2.5.3). Since  $(x_j, g(x_j)) \in \text{epi } g$  for  $j = 1, \dots, k$ ,

$$\left( x, \sum_{j=1}^k \alpha_j g(x_j) \right) \in \text{co epi } g$$

and this implies

$$f_1(x) \leq \sum_{j=1}^k \alpha_j g(x_j)$$

by definition of  $f_1$ . Because the decomposition of  $x$  was arbitrary within (2.5.3), this implies  $f_1(x) \leq f_3(x)$ .  $\square$

Note in (2.5.3) the role of the convention  $\inf \emptyset = +\infty$ , in case  $x$  has no decomposition – which means that  $x \notin \text{co dom } g$ . The restrictions  $x_j \in \text{dom } g$  could be equally relaxed (an  $x_j \notin \text{dom } g$  certainly does not help making the infimum); notationally,  $\alpha$  should then be taken in  $\text{ri } \Delta_k$ , so as to avoid the annoying multiplication  $0 \times (+\infty)$ . Beware that  $\text{epi}(\text{co } g)$  is *not exactly* the convex hull  $\text{co}(\text{epi } g)$ : we need to close the bottom of this latter set, as in §1.3(g)(ii) – an operation which affects only the relative boundary of  $\text{co epi } g$ , though. Note also that Carathéodory's Theorem yields an upper bound on  $k$  for (2.5.3), namely  $k \leq (n+1)+1 = n+2$ . We just mention a property which is of little use for the time being: the upper bound can be reduced to  $k \leq n+1$  (see Proposition III.4.2.3).

Instead of  $\text{co epi } g$ , we can take the closed convex hull  $\bar{\text{co}} \text{ epi } g = \text{cl co epi } g$  (see §III.1.4). We obtain a closed set, with in particular a closed bottom: it is already an epigraph, the epigraph of a closed convex function. The corresponding operation that yielded  $f_1, f_2, f_3$  is therefore now simpler. Furthermore, we know from Proposition 1.2.8 that all closed convex functions are redundant to define the function corresponding to  $f_2$ : affine functions are enough. We leave it as an exercise to prove the following result:

**Proposition 2.5.2** *Let  $g$  satisfy the hypotheses of Proposition 2.5.1. Then the three functions below*

$$\begin{aligned}\bar{f}_1(x) &:= \inf \{r : (x, r) \in \bar{\text{co}} \text{ epi } g\}, \\ \bar{f}_2(x) &:= \sup \{h(x) : h \in \text{Conv } \mathbb{R}^n, h \leq g\}, \\ \bar{f}_3(x) &:= \sup \{\langle s, x \rangle - b : \langle s, y \rangle - b \leq g(y) \text{ for all } y \in \mathbb{R}^n\}\end{aligned}$$

*are closed, convex, and coincide on  $\mathbb{R}^n$  with the closure of the function constructed in Proposition 2.5.1.*  $\square$

In view of the relationship between the operations studied in this Section 2.5 and the convexification of  $\text{epi } g$ , the following notation is justified, even if it is not quite accurate.

**Definition 2.5.3 (Convex Hulls of a Function)** Let  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n \cup \{+\infty\}$ , not identically  $+\infty$ , be minorized by an affine function. The common function  $f_1 = f_2 = f_3$  of Proposition 2.5.1 is called the *convex hull* of  $g$ , denoted  $\text{co } g$ . The *closed convex hull* of  $g$  is any of the functions described by Proposition 2.5.2; it is denoted  $\bar{\text{co}} g$  or  $\text{cl co } g$ .  $\square$

If  $\{g_j\}_{j \in J}$  is an arbitrary family of functions, all minorized by the same affine function, the epigraph of the [closed] convex hull of the function  $\inf_{j \in J} g_j$  is obtained from  $\cup_{j \in J} \text{epi } g_j$ . An important case is when the  $g_j$ 's are *convex*; then, exploiting Example III.1.3.5, the formula giving  $f_3$  simplifies: several  $x_j$ 's corresponding to the same  $g_i$  can be compressed to a single convex combination.

**Proposition 2.5.4** Let  $g_1, \dots, g_m$  be in  $\text{Conv } \mathbb{R}^n$ , all minorized by the same affine function. Then the convex hull of their infimum is defined by

$$\begin{aligned} \mathbb{R}^n &\ni x \mapsto [\text{co}(\min_j g_j)](x) = \\ \inf \left\{ \sum_{j=1}^m \alpha_j g_j(x_j) : \alpha \in \Delta_m, x_j \in \text{dom } g_j, \sum_{j=1}^m \alpha_j x_j = x \right\}. \end{aligned} \quad (2.5.4)$$

PROOF. Apply Example III.1.3.5 to the convex sets  $C_j = \text{epi } g_j$ . □

The above statement was made in the simple situation of finitely many  $g_j$ 's, but the representation (2.5.4) can be extended to an arbitrary family of convex functions  $g_j$ : it suffices to consider in the infimand all the representations of  $x$  as convex combinations of finitely many elements  $x_j \in \text{dom } g_j$ .

Along these lines, note that an arbitrary function  $g : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  can be seen as an infimum of convex functions: considering  $\text{dom } g$  as an index-set,

$$g(x) = \inf \left\{ g(x_j) + I_{\{x_j\}}(x) : x_j \in \text{dom } g \right\},$$

where each  $g(x_j)$  denotes a (finite) constant function.

**Example 2.5.5** Let  $(x_1, b_1), \dots, (x_m, b_m)$  be given in  $\mathbb{R}^n \times \mathbb{R}$  and define for  $j = 1, \dots, m$

$$g_j(x) = \begin{cases} b_j & \text{if } x = x_j, \\ +\infty & \text{if not.} \end{cases}$$

Then  $f := \text{co}(\min g_j) = \bar{\text{co}}(\min g_j)$  is the polyhedral function with the epigraph illustrated on Fig. 2.5.1, and analytically given by

$$f(x) = \begin{cases} \min \left\{ \sum_{j=1}^m \alpha_j b_j : \alpha \in \Delta_m, \sum_{j=1}^m \alpha_j x_j = x \right\} & \text{if } x \in \text{co}\{x_1, \dots, x_m\}, \\ +\infty & \text{if not.} \end{cases}$$

Calling  $b \in \mathbb{R}^m$  the vector whose components are the  $b_i$ 's and  $A$  the matrix whose columns are the  $x_i$ 's, the above minimization problem in  $\alpha$  can be written – at least when  $x \in \text{co}\{x_1, \dots, x_m\}$ :

$$f(x) = \min \{b^\top \alpha : \alpha \in \Delta_m, A\alpha = x\}.$$
□

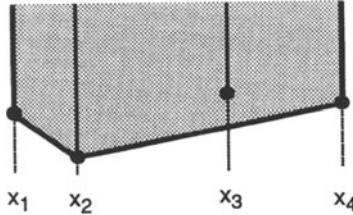


Fig. 2.5.1. A convex hull of needles

To conclude this Section 2, Table 2.5.1 summarizes the main operations on functions and epigraphs that we have encountered.

**Table 2.5.1.** Main operations yielding convexity

Operations on functions: $f =$	Operations on sets: $\text{epi } f$ or $\text{epi}_s f =$	Closedness
$\sum_{j=1}^m t_j f_j$	nothing interesting	preserved
$\sup_{j \in J} f_j$	$\cap_{j \in J} \text{epi } f_j$	preserved
$g \circ A$ ( $A$ affine)	$A'(\text{epi } g)$	preserved
$ug(x/u)$	$\mathbb{R}_*^+(\{1\} \times \text{epi } g)$	must be forced
$f_1 \downarrow f_2$	$\text{epi}_s f_1 + \text{epi}_s f_2$	destroyed
$f_1 \bar{\vee} f_2$	$\text{epi}_s f_1 \pm \text{epi}_s f_2$	preserved from $f_1$
$Ag$ ( $A$ linear)	epigr. hull of $A'(\text{epi } g)$	destroyed
$\inf_y g(\cdot, y)$	$\text{Proj}_{\mathbb{R}^n \times \mathbb{R}} \text{epi}_s g$	destroyed
$\text{co } g$	epigr. hull of $\text{co epi } g$	can be forced

### 3 Local and Global Behaviour of a Convex Function

#### 3.1 Continuity Properties

Convex functions turn out to enjoy remarkable continuity properties: as already seen in §I.3, they are locally Lipschitzian on the relative interior of their domain. On the relative boundary of that domain, however, all kinds of continuity may disappear.

We start with a technical lemma.

**Lemma 3.1.1** *Let  $f \in \text{Conv } \mathbb{R}^n$  and suppose there are  $x_0, \delta, m$  and  $M$  such that*

$$m \leq f(x) \leq M \quad \text{for all } x \in B(x_0, 2\delta).$$

*Then  $f$  is Lipschitzian on  $B(x_0, \delta)$ ; more precisely: for all  $y$  and  $y'$  in  $B(x_0, \delta)$ ,*

$$|f(y) - f(y')| \leq \frac{M - m}{\delta} \|y - y'\|. \quad (3.1.1)$$

PROOF. Look at Fig. 3.1.1: with two different  $y$  and  $y'$  in  $B(x_0, \delta)$ , take

$$y'' := y' + \delta \frac{y' - y}{\|y' - y\|} \in B(x_0, 2\delta);$$

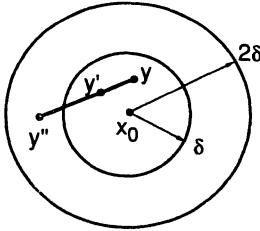
by construction,  $y'$  lies on the segment  $[y, y'']$ , namely

$$y' = \frac{\|y' - y\|}{\delta + \|y' - y\|} y'' + \frac{\delta}{\delta + \|y' - y\|} y.$$

Applying the convexity of  $f$  and using the postulated bounds, we obtain

$$f(y') - f(y) \leq \frac{\|y' - y\|}{\delta + \|y' - y\|} [f(y'') - f(y)] \leq \frac{1}{\delta} \|y' - y\| (M - m).$$

Then, it suffices to exchange  $y$  and  $y'$  to prove (3.1.1).  $\square$



**Fig. 3.1.1.** Moving in a neighborhood of  $x_0$

**Theorem 3.1.2** *With  $f \in \text{Conv } \mathbb{R}^n$ , let  $S$  be a convex compact subset of  $\text{ri dom } f$ . Then there exists  $L = L(S) \geq 0$  such that*

$$|f(x) - f(x')| \leq L \|x - x'\| \quad \text{for all } x \text{ and } x' \text{ in } S. \quad (3.1.2)$$

**PROOF. [Preliminaries]** First of all, our statement ignores  $x$ -values outside the affine hull of the convex set  $\text{dom } f$ . Instead of  $\mathbb{R}^n$ , it can be formulated in  $\mathbb{R}^d$ , where  $d$  is the dimension of  $\text{dom } f$ ; alternatively, we may assume  $\text{ri dom } f = \text{int dom } f$ , which will simplify the writing.

Make this assumption and let  $x_0 \in S$ . We will prove that there are  $\delta = \delta(x_0) > 0$  and  $L = L(x_0, \delta)$  such that the ball  $B(x_0, \delta)$  is included in  $\text{int dom } f$  and

$$|f(y) - f(y')| \leq L \|y - y'\| \quad \text{for all } y \text{ and } y' \text{ in } B(x_0, \delta). \quad (3.1.3)$$

If this holds for all  $x_0 \in S$ , the corresponding balls  $B(x_0, \delta)$  will provide a covering of the compact set  $S$ , from which we will extract a finite covering  $(x_1, \delta_1, L_1), \dots, (x_k, \delta_k, L_k)$ . With these balls, we will divide an arbitrary segment  $[x, x']$  of the convex set  $S$  into finitely many subsegments, of endpoints  $y_0 := x, \dots, y_i, \dots, y_\ell := x'$ . Ordering properly the  $y_i$ 's, we will have  $\|x - x'\| = \sum_{i=1}^\ell \|y_i - y_{i-1}\|$ ; furthermore,  $f$  will be Lipschitzian on each  $[y_{i-1}, y_i]$  with the common constant  $L := \max\{L_1, \dots, L_k\}$ . The required Lipschitz property (3.1.2) will follow.

**[Main Step]** To establish (3.1.3), we use Lemma 3.1.1, which requires boundedness of  $f$  in the neighborhood of  $x_0$ . For this, we construct as in the proof of Theorem III.2.1.3 (see Fig. III.2.1.1) a simplex

$$\Delta = \text{co}\{v_0, \dots, v_n\} \subset \text{dom } f$$

having  $x_0$  in its interior: we can take  $\delta > 0$  such that  $B(x_0, 2\delta) \subset \Delta$ . Then any  $y \in B(x_0, 2\delta)$  can be written – we use the notation (2.5.1):

$$y = \sum_{i=0}^n \alpha_i v_i \quad \text{with} \quad \alpha \in \Delta_{n+1},$$

so that the convexity of  $f$  gives

$$f(y) \leq \sum_{i=0}^n \alpha_i f(v_i).$$

On the other hand, Proposition 1.2.1 tells us that  $f$  is bounded from below, say by  $m$ , on this very same  $B(x_0, 2\delta)$ . Our claim is proved: we have singled out  $\delta > 0$  such that, with  $M := \max\{f(v_0), \dots, f(v_n)\}$ ,

$$m \leq f(y) \leq M \quad \text{for all } y \in B(x_0, 2\delta). \quad \square$$

Note that the key-argument in the main step above is to find a (relative) neighborhood of  $x \in \text{ri dom } f$ , which is convex and which has a finite number of extreme points, all lying in  $\text{dom } f$ . The simplex  $\Delta$  is such a neighborhood, with a minimal number of extreme points.

**Remark 3.1.3** It follows in particular that  $f$  is continuous relatively to the relative interior of its domain, i.e.: for  $x_0 \in \text{ri dom } f$  and  $x \in \text{ri dom } f$  converging to  $x_0$ , we have that  $f(x) \rightarrow f(x_0)$ .

An equivalent formulation of Theorem 3.1.2 is:  $f$  is locally Lipschitzian on the relative interior of its domain, i.e. for all  $x_0 \in \text{ri dom } f$ , there are  $L(x_0)$  and  $\delta(x_0)$  such that

$$\begin{aligned} |f(x) - f(x')| &\leq L(x_0)\|x - x'\| \quad \text{for all } x \text{ and } x' \text{ in the set} \\ S(x_0) &:= B(x_0, \delta(x_0)) \cap \text{aff dom } f \subset \text{ri dom } f. \end{aligned}$$

In fact, the bulk of our proof is just concerned with this last statement. Of course, when  $x_0$  gets closer to the relative boundary of  $\text{dom } f$ , the size  $\delta(x_0)$  of the allowed neighborhood shrinks to 0; but also, the local Lipschitz constant  $L(x_0)$  may grow unboundedly ( $g_f$  may become steeper and steeper).  $\square$

Because of the phenomenon mentioned in the above remark, we cannot put  $\text{ri dom } f$  instead of  $S$  in Theorem 3.1.2: a convex function need not be Lipschitzian on the relative interior of its domain. However, it is possible to modify  $f$  outside the given compact  $S$ , and to obtain a convex function which is Lipschitzian on the whole space:

**Proposition 3.1.4 (Lipschitzian Extension)** *Let  $C$  be a nonempty convex set, and let  $f \in \text{Conv } \mathbb{R}^n$  be Lipschitzian with constant  $L$  on  $C$ . Then there exists a convex function  $f_1$  satisfying*

$$f_1(x) = f(x) \quad \text{for all } x \in C, \quad (3.1.4)$$

$$f_1 \text{ is Lipschitzian with constant } L \text{ on the whole space.} \quad (3.1.5)$$

Moreover, there is a largest function satisfying (3.1.4), (3.1.5), namely the infimal convolution

$$\begin{aligned} \mathbb{R}^n \ni x \mapsto (f + I_C)_{[L]}(x) &:= [(f + I_C) \downarrow (L\|\cdot\|)](x) \\ &= \inf \{f(y) + L\|x - y\| : y \in C\}. \end{aligned} \quad (3.1.6)$$

PROOF. Call  $\hat{f}$  the function (3.1.6). First we show that  $\hat{f}(x) > -\infty$  for all  $x$ . In fact, let  $x_0 \in \text{ri } C$  and apply Proposition 1.2.1 to the function  $f + I_C$ , whose domain is clearly  $C$ : there is  $s$  in the subspace  $V$  parallel to  $\text{aff } C$  such that

$$f(x) \geq f(x_0) + \langle s, x - x_0 \rangle \quad \text{for all } x \in \mathbb{R}^n.$$

Taking  $\delta > 0$  so small that  $x = x_0 + \delta s$  is in  $C$ , we obtain with the Lipschitz property of  $f$  on  $C$ :

$$L\delta\|s\| \geq f(x) - f(x_0) \geq \delta\|s\|^2,$$

whence  $\|s\| \leq L$ . Then

$$\langle s, x \rangle \leq \|s\| \|x\| \leq L\|x\| \quad \text{for all } x \in \mathbb{R}^n.$$

Thus, the two functions making up  $\hat{f}$  are minorized by a common affine function (with slope  $s$ ): in view of Proposition 2.3.2,  $\hat{f} \in \text{Conv } \mathbb{R}^n$ .

Next, given  $x$  and  $x'$  in  $\mathbb{R}^n$  and  $\varepsilon > 0$ , let  $y' \in C$  be such that

$$f(y') + L\|x' - y'\| \leq \hat{f}(x') + \varepsilon;$$

by definition, we also have

$$\hat{f}(x) \leq f(y') + L\|x - y'\| \leq f(y') + L\|x - x'\| + L\|x' - y'\|,$$

so we obtain

$$\hat{f}(x) \leq \hat{f}(x') + L\|x - x'\| + \varepsilon.$$

This relation holds for arbitrary  $x, x'$  and  $\varepsilon$ , so it does imply the Lipschitz property of  $\hat{f}$  on  $\mathbb{R}^n$ .

Now let  $x \in C$ . Again by definition,  $\hat{f}(x) \leq f(x)$ ; and also, the Lipschitz property of  $f$  on  $C$  implies

$$f(x) \leq f(y) + L\|y - x\| \quad \text{for all } y \in C,$$

so  $f(x) \leq \hat{f}(x)$ . In a word,  $\hat{f}$  coincides with  $f$  on  $C$ .

Finally, let  $f_1$  satisfy (3.1.4), (3.1.5). We obtain in particular

$$f_1(x) - f(y) \leq L\|x - y\| \quad \text{for all } x \in \mathbb{R}^n \text{ and } y \in C,$$

so  $f_1$  minorizes  $\hat{f}$  on  $\mathbb{R}^n$  and the proof is complete.  $\square$

Constructing from the given  $f$  and  $C$  the Lipschitzian function of (3.1.6) thus appears as a sort of regularization. Such a mechanism is often useful and will be encountered again.

Let us sum up the continuity properties of a convex function.

- First of all it is  $\text{aff dom } f$ , and not  $\mathbb{R}^n$ , that is the relevant embedding (affine) space: there is no point in studying the behaviour of  $f$  when moving out of this space. Continuity, and even Lipschitz continuity, holds when  $x$  remains “well inside”  $\text{ri dom } f$ .
- When  $x$  approaches  $\text{rbd dom } f$ , continuity may break down:  $f$  may go to infinity, or jump discontinuously to some finite value, etc. Still, irregular behaviour of  $f$  is limited by Proposition 1.2.5.
- Closing  $\text{epi } f$  if necessary, lower semi-continuity of  $f$  is a tolerable assumption. Doing this, we only miss functions having little interest in our framework of minimization.

- It remains to ask whether  $f$  can be assumed upper semi-continuous (on  $\text{rbd dom } f$ , and relative to  $\text{dom } f$ ): we have seen in §I.3.1 that this property automatically holds for univariate functions. The answer is no in general, though: a counter-example is

$$\mathbb{R}^2 \ni x = (\xi, \eta) \mapsto f(x) = \sup_{\alpha, \beta} \{ \xi\alpha + \eta\beta : \frac{1}{2}\alpha^2 \leq \beta \} .$$

We see that  $f(0) = 0$ , and we know from Proposition 2.1.2 that  $f \in \text{Conv} \mathbb{R}^n$ . In fact, the optimal  $(\alpha, \beta)$  (if any) satisfies  $\frac{1}{2}\alpha^2 = \beta$ , so that

$$f(\xi, \eta) = \sup_{\alpha} \left( \frac{1}{2}\eta\alpha^2 + \xi\alpha \right) = \begin{cases} 0 & \text{if } \xi = \eta = 0, \\ -\frac{\xi^2}{2\eta} & \text{if } \eta < 0, \\ +\infty & \text{otherwise.} \end{cases} \quad (3.1.7)$$

Thus, when  $x$  tends to 0 following the path  $\eta = -\frac{1}{2}\xi^2$ , then  $f(x) \equiv 1 > 0 = f(0)$ .

To conclude this subsection, we give a rather powerful convergence result: convex functions converging pointwise to some (convex) function  $f$  do converge *uniformly* on each compact set contained in the relative interior of  $\text{dom } f$ . For the sake of simplicity, we limit ourselves here to the case of finite-valued functions. For the general case, just specify that the compact set  $S$  in the next statement must be in  $\text{ri dom } f$ , and adapt the proof accordingly.

**Theorem 3.1.5** *Let the convex functions  $f_k : \mathbb{R}^n \rightarrow \mathbb{R}$  converge pointwise for  $k \rightarrow +\infty$  to  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . Then  $f$  is convex and, for each compact set  $S$ , the convergence of  $f_k$  to  $f$  is uniform on  $S$ .*

PROOF. Convexity of  $f$  is trivial: pass to the limit in the definition (1.1.1) itself. For uniformity, we want to use Lemma 3.1.1, so we need to bound  $f_k$  on  $S$  independently of  $k$ ; thus, let  $r > 0$  be such that  $S \subset B(0, r)$ .

[Step 1] First the function  $g := \sup_k f_k$  is convex, and  $g(x) < +\infty$  for all  $x$  because the convergent sequence  $\{f_k(x)\}$  is certainly bounded. Hence,  $g$  is continuous and therefore bounded, say by  $M$ , on the compact set  $B(0, 2r)$ :

$$f_k(x) \leq g(x) \leq M \quad \text{for all } k \text{ and all } x \in B(0, 2r).$$

Second, the convergent sequence  $\{f_k(0)\}$  is bounded from below:

$$\mu \leq f_k(0) \quad \text{for all } k.$$

Then, for  $x \in B(0, 2r)$  and all  $k$ , write the convexity relation on  $[-x, x] \subset B(0, 2r)$ :

$$2\mu \leq 2f_k(0) \leq f_k(x) + f_k(-x) \leq f_k(x) + M,$$

i.e. the  $f_k$ 's are bounded from below, independently of  $k$ . Thus, we are within the conditions of Lemma 3.1.1: there is some  $L$  (independent of  $k$ ) such that

$$|f_k(y) - f_k(y')| \leq L\|y - y'\| \quad \text{for all } k \text{ and all } y, y' \text{ in } B(0, r). \quad (3.1.8)$$

Naturally, the same Lipschitz property is transmitted to the limiting function  $f$ .

[Step 2] Now fix  $\varepsilon > 0$ . Cover  $S$  by the balls  $B(x, \varepsilon)$  for  $x$  describing  $S$ , and extract a finite covering  $S \subset B(x_1, \varepsilon) \cup \dots \cup B(x_m, \varepsilon)$ . With  $x$  arbitrary in  $S$ , take an  $x_i$  such that  $x \in B(x_i, \varepsilon)$ . There is  $k_{i,\varepsilon}$  such that, for all  $k \geq k_{i,\varepsilon}$ ,

$$|f_k(x) - f(x)| \leq |f_k(x) - f_k(x_i)| + |f_k(x_i) - f(x_i)| + |f(x_i) - f(x)| \leq (2L + 1)\varepsilon$$

where we have also used (3.1.8), knowing that  $x$  and  $x_i$  are in  $S \subset B(0, r)$ . The above inequality is then valid uniformly in  $x$ , providing that

$$k \geq \max\{k_{1,\varepsilon}, \dots, k_{m,\varepsilon}\} =: k_\varepsilon.$$

□

### 3.2 Behaviour at Infinity

Having studied the behaviour of  $f(x)$  when  $x$  approaches  $\text{rbdom } f$ , it remains to consider the case of unbounded  $x$ . An important issue is the behaviour of  $f(x_0 + td)$  when  $t \rightarrow +\infty$  ( $x_0$  and  $d$  being fixed). It has already been addressed in the simple situation of §I.2.3, where only two directions  $d = \pm 1$  had to be considered. Here we have infinitely many directions, but  $\text{epi } f$  is after all a special unbounded convex set of  $\mathbb{R}^{n+1}$ ; so we can use the results of §III.2.2.

Thus we assume  $f \in \text{Conv} \mathbb{R}^n$ , which allows us to consider the asymptotic cone  $(\text{epi } f)_\infty$  of the closed convex set  $\text{epi } f$ . It is a closed convex cone of  $\mathbb{R}^n \times \mathbb{R}$ , which clearly contains the half-line  $\{0\} \times \mathbb{R}^+$ . According to its Definition III.2.2.2,

$$(\text{epi } f)_\infty = \{(d, \rho) \in \mathbb{R}^n \times \mathbb{R} : (x_0, r_0) + t(d, \rho) \in \text{epi } f \text{ for all } t > 0\}, \quad (3.2.1)$$

where  $(x_0, r_0)$  is an arbitrary element of  $\text{epi } f$ . This can be written

$$(\text{epi } f)_\infty = \{(d, \rho) : \text{epi } f + t(d, \rho) \subset \text{epi } f \text{ for all } t > 0\}$$

and, since we already know that  $(\text{epi } f)_\infty$  is a convex cone:

$$(\text{epi } f)_\infty = \{(d, \rho) : \text{epi } f + (d, \rho) \subset \text{epi } f\}.$$

**Remark 3.2.1** Such an object was already encountered in Example III.1.2.6: we are dealing with the star-difference between  $\text{epi } f$  and itself:

$$(\text{epi } f)_\infty = \text{epi } f * \text{epi } f.$$

This in turn was seen in Remark 2.3.9, and it shows that  $(\text{epi } f)_\infty$  is itself an epigraph: the epigraph of the deconvolution of  $f$  by itself:

$$(\text{epi } f)_\infty = \text{epi}(f \bar{\vee} f).$$

In other words, the behaviour of  $f$  at infinity can be described with the help of the function

$$(f \bar{\vee} f)(d) = \sup \{f(x + d) - f(x) : x \in \text{dom } f\}. \quad (3.2.2)$$

□

Using directly the definition (3.2.1) of  $(\text{epi } f)_\infty$ , there is an alternate way of expressing the function (3.2.2):

**Proposition 3.2.2** For  $f \in \text{Conv} \mathbb{R}^n$ , the asymptotic cone of  $\text{epi } f$  is the epigraph of the function  $f'_\infty \in \text{Conv} \mathbb{R}^n$  defined by

$$d \mapsto f'_\infty(d) := \sup_{t>0} \frac{f(x_0 + td) - f(x_0)}{t} = \lim_{t \rightarrow +\infty} \frac{f(x_0 + td) - f(x_0)}{t}, \quad (3.2.3)$$

where  $x_0$  is arbitrary in  $\text{dom } f$ .

PROOF. Since  $(x_0, f(x_0))$  is an element of  $\text{epi } f$ , (3.2.1) tells us that  $(d, \rho) \in (\text{epi } f)_\infty$  if and only if

$$f(x_0 + td) \leq f(x_0) + t\rho \quad \text{for all } t > 0,$$

which means

$$\sup_{t>0} \frac{f(x_0 + td) - f(x_0)}{t} \leq \rho. \quad (3.2.4)$$

In other words,  $(\text{epi } f)_\infty$  is the epigraph of the function whose value at  $d$  is the left-hand side of (3.2.4); and this is true no matter how  $x_0$  has been chosen in  $\text{dom } f$ . The rest follows from the fact that the difference quotient in (3.2.4) is closed convex in  $d$ , and increasing in  $t$  (the function  $t \mapsto f(x_0 + td)$  is convex and enjoys the property of increasing slopes, remember Proposition I.1.1.4).  $\square$

It goes without saying that the expressions appearing in (3.2.3) are independent of  $x_0$ :  $f'_\infty$  is really a function of  $d$  only. By construction, this function is positively homogeneous:

$$f'_\infty(\alpha d) = \alpha f'_\infty(d) \quad \text{for all } \alpha > 0.$$

Our notation suggests that it is something like a “slope at infinity” in the direction  $d$ .

**Definition 3.2.3** The function  $f'_\infty$  of Proposition 3.2.2 is called the *asymptotic function*, or recession function, or auto-deconvolution, of  $f$ .  $\square$

Consider for example the indicator  $I_C$  of a closed convex set  $C$ . By definition of the asymptotic cone, we see that  $I_C(x_0 + td) = 0$  for all  $t > 0$  if and only if  $d \in C_\infty$ ; we obtain

$$(I_C)'_\infty = I_{(C_\infty)}.$$

The next example is more interesting and extends Remark 2.2.3:

**Example 3.2.4** Let  $f \in \text{Conv} \mathbb{R}^n$ . Take  $x_0 \in \text{dom } f$  and consider the convex function  $d \mapsto f(x_0 + d) - f(x_0)$ , whose domain contains 0, and whose perspective-function is  $r$  of (2.2.2). The closure of  $r$  can be computed with the help of Proposition 2.2.2: with  $x_0 + d'$  arbitrary in  $\text{ri dom } f$ ,

$$(\text{cl } r)(0, d) = \lim_{\alpha \downarrow 0} \alpha[f(x_0 + d' - d + d/\alpha) - f(x_0)].$$

Note that the term  $f(x_0) < +\infty$  can be suppressed, or replaced by  $f(x_0 + d')$  (because  $\alpha \downarrow 0$ ); moreover, as in Remark 2.2.3, the above limit is exactly

$$\lim_{t \rightarrow +\infty} \frac{f(x_0 + d' + td)}{t} = \lim_{t \rightarrow +\infty} \frac{f(x_0 + d' + td) - f(x_0 + d')}{t} = f'_\infty(d').$$

In summary, the function defined by

$$\mathbb{R} \times \mathbb{R}^n \ni (u, d) \mapsto \begin{cases} u[f(x_0 + d/u) - f(x_0)] & \text{if } u > 0, \\ f'_\infty(d) & \text{if } u = 0, \\ +\infty & \text{elsewhere} \end{cases}$$

is in  $\overline{\text{Conv}}(\mathbb{R} \times \mathbb{R}^n)$ ; and only its “ $u > 0$ -part” depends on the reference point  $x_0 \in \text{dom } f$ .  $\square$

Our next result assesses the importance of the asymptotic function.

**Proposition 3.2.5** *Let  $f \in \overline{\text{Conv}} \mathbb{R}^n$ . All the nonempty sublevel-sets of  $f$  have the same asymptotic cone, which is the sublevel-set of  $f'_\infty$  at the level 0:*

$$\forall r \in \mathbb{R} \text{ with } S_r(f) \neq \emptyset, \quad [S_r(f)]_\infty = \{d \in \mathbb{R}^n : f'_\infty(d) \leq 0\}.$$

In particular, the following statements are equivalent:

- (i) There is  $r$  for which  $S_r(f)$  is nonempty and compact;
- (ii) all the sublevel-sets of  $f$  are compact;
- (iii)  $f'_\infty(d) > 0$  for all nonzero  $d \in \mathbb{R}^n$ .

PROOF. By definition (III.2.2.1), a direction  $d$  is in the asymptotic cone of the nonempty sublevel-set  $S_r(f)$  if and only if

$$x \in S_r(f) \implies [x + td \in S_r(f) \text{ for all } t > 0],$$

which can also be written – see (1.1.4):

$$(x, r) \in \text{epi } f \implies (x + td, r + t \times 0) \in \text{epi } f \text{ for all } t > 0;$$

and this in turn just means that  $(d, 0) \in (\text{epi } f)_\infty = \text{epi } f'_\infty$ . We have proved the first part of the theorem.

A particular case is when the sublevel-set  $S_0(f'_\infty)$  is reduced to the singleton  $\{0\}$ , which exactly means (iii). This is therefore equivalent to

$$[S_r(f)]_\infty = \{0\} \text{ for all } r \in \mathbb{R} \text{ with } S_r(f) \neq \emptyset,$$

which means that  $S_r(f)$  is compact (Proposition III.2.2.3). The equivalence between (i), (ii) and (iii) is proved.  $\square$

Needless to say, the convexity of  $f$  is essential to ensure that all its nonempty sublevel-sets have the same asymptotic cone. In Remark 1.1.7, we have seen (closed) quasi-convex functions: their sublevel-sets are all convex, and as such they have asymptotic cones, which normally depend on the level.

**Definition 3.2.6 (Coercivity)** The functions  $f \in \overline{\text{Conv}} \mathbb{R}^n$  satisfying (i), (ii) or (iii) are called *0-coercive*. Equivalently, the 0-coercive functions are those that “increase at infinity”:

$$f(x) \rightarrow +\infty \text{ whenever } \|x\| \rightarrow +\infty,$$

and closed convex 0-coercive functions achieve their minimum over  $\mathbb{R}^n$ .

An important particular case is when  $f'_\infty(d) = +\infty$  for all  $d \neq 0$ , i.e.  $f'_\infty = I_{\{0\}}$ . It can be seen that this means precisely

$$\frac{f(x)}{\|x\|} \rightarrow +\infty \quad \text{whenever} \quad \|x\| \rightarrow +\infty.$$

In words: at infinity,  $f$  increases to infinity faster than any affine function (to establish this equivalence, extract a cluster point of  $\{x_k/\|x_k\|\}$  and use the lower semi-continuity of  $f'_\infty$ ). Such functions are called *1-coercive*, or sometimes just coercive.  $\square$

Suppose for example that  $f$  is quadratic:

$$f(x) = \frac{1}{2}\langle Qx, x \rangle + \langle b, x \rangle + c,$$

with  $Q$  a positive semi-definite symmetric operator,  $b \in \mathbb{R}^n$  and  $c \in \mathbb{R}$ . Then it is easy to compute

$$f'_\infty(d) = \begin{cases} \langle b, d \rangle & \text{if } d \in \text{Ker } Q, \\ +\infty & \text{if not.} \end{cases}$$

In this particular case, the different sorts of coercivity coincide:

$$f \text{ is 0-coercive} \iff f \text{ is 1-coercive} \iff Q \text{ is positive definite.}$$

The word “coercive” alone comes from the study of bilinear forms: for our more general framework of non-quadratic functions, it becomes ambiguous, hence our distinction.

**Proposition 3.2.7** *A function  $f \in \text{Conv} \mathbb{R}^n$  is Lipschitzian on the whole of  $\mathbb{R}^n$  if and only if  $f'_\infty$  is finite on the whole of  $\mathbb{R}^n$ . The best Lipschitz constant for  $f$  is then*

$$\sup \{f'_\infty(d) : \|d\| = 1\}. \quad (3.2.5)$$

PROOF. When the (convex) function  $f'_\infty$  is finite-valued, it is continuous (§3.1) and therefore bounded on the compact unit sphere:

$$\sup \{f'_\infty(d) : \|d\| = 1\} =: L < +\infty,$$

which implies by positive homogeneity

$$f'_\infty(d) \leq L\|d\| \quad \text{for all } d \in \mathbb{R}^n.$$

Now use the definition (3.2.2) of  $f'_\infty$ :

$$f(x + d) - f(x) \leq L\|d\| \quad \text{for all } x \in \text{dom } f \text{ and } d \in \mathbb{R}^n;$$

thus,  $\text{dom } f$  is the whole space ( $f(x + d) < +\infty$  for all  $d$ ) and we do obtain that  $L$  is a global Lipschitz constant for  $f$ .

Conversely, let  $f$  have a global Lipschitz constant  $L$ . Pick  $x_0 \in \text{dom } f$  and plug the inequality

$$f(x_0 + td) - f(x_0) \leq Lt\|d\| \quad \text{for all } t > 0 \text{ and } d \in \mathbb{R}^n$$

into the definition (3.2.3) of  $f'_\infty$  to obtain

$$f'_\infty(d) \leq L\|d\| \quad \text{for all } d \in \mathbb{R}^n.$$

It follows that  $f'_\infty$  is finite everywhere, and the value (3.2.5) does not exceed  $L$ .  $\square$

Concerning (3.2.5), it is worth mentioning that the index-set  $\tilde{B}$  can be replaced by the unit ball  $B$ , and/or absolute value can be inserted in the supremand; these two replacements are made possible thanks to convexity.

**Remark 3.2.8** We mention the following classification, of interest for minimization theory: let  $f \in \text{Conv} \mathbb{R}^n$ .

- If  $f'_\infty(d) < 0$  for some  $d$ , then  $f$  is unbounded from below; more precisely: for all  $x_0 \in \text{dom } f$ ,  $f(x_0 + td) \downarrow -\infty$  when  $t \rightarrow +\infty$ .
- The condition  $f'_\infty(d) > 0$  for all  $d \neq 0$  is necessary and sufficient for  $f$  to have a nonempty bounded (hence compact) set of minimum points.
- If  $f'_\infty \geq 0$ , with  $f'_\infty(d) = 0$  for some  $d \neq 0$ , existence of a minimum cannot be guaranteed (but if  $x_0$  is minimal, so is the half-line  $x_0 + \mathbb{R}^+ d$ ).

Observe that, if the continuous function  $d \mapsto f'_\infty(d)$  is positive for all  $d \neq 0$ , then it is minorized by some  $m > 0$  on the unit sphere  $\tilde{B}$  and this  $m$  also minorizes the speed at which  $f$  increases at infinity.  $\square$

To close this section, we mention some calculus rules on the asymptotic function. They come directly either from the analytical definitions (3.2.2), (3.2.3), or from the geometrical definition  $\text{epi } f'_\infty = (\text{epi } f)_\infty$  combined with Proposition III.2.2.5.

### Proposition 3.2.9

- Let  $f_1, \dots, f_m$  be  $m$  functions of  $\text{Conv} \mathbb{R}^n$ , and  $t_1, \dots, t_m$  be positive numbers. Assume that there is  $x_0$  at which each  $f_j$  is finite. Then, for  $f := \sum_{j=1}^m t_j f_j$ ,

$$f'_\infty = \sum_{j=1}^m t_j (f_j)'_\infty.$$

- Let  $\{f_j\}_{j \in J}$  be a family of functions in  $\text{Conv} \mathbb{R}^n$ . Assume that there is  $x_0$  at which  $\sup_{j \in J} f_j(x_0) < +\infty$ . Then, for  $f := \sup_{j \in J} f_j$ ,

$$f'_\infty = \sup_{j \in J} (f_j)'_\infty.$$

- Let  $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$  be affine with linear part  $A_0$ , and let  $f \in \text{Conv} \mathbb{R}^m$ . Assume that  $A(\mathbb{R}^n) \cap \text{dom } f \neq \emptyset$ . Then

$$(f \circ A)'_\infty = f'_\infty \circ A_0.$$

$\square$

For an image-function (2.4.2), the corresponding formula is

$$(Ag)'_\infty(d) = \inf \{g'_\infty(z) : Az = d\}$$

which can be written symbolically:  $(Ag)'_\infty = A(g'_\infty)$ . However, this formula cannot hold without an additional assumption, albeit to guarantee  $Ag \in \text{Conv } \mathbb{R}^n$ . One such assumption is

$$g'_\infty(z) > 0 \quad \text{for all } z \in \text{Ker } A \setminus \{0\}$$

(appropriate coercivity is added where necessary, so that the infimum in the definition of  $(Ag)(x)$  is always attained). Proving this result is not simple.

## 4 First- and Second-Order Differentiation

Let  $C \subset \mathbb{R}^n$  be nonempty and convex. For a function  $f$  defined on  $C$  ( $f(x) < +\infty$  for all  $x \in C$ ), we study here the following questions:

- When  $f$  is convex and differentiable on  $C$ , what can be said about the gradient  $\nabla f$ ?
- When  $f$  is differentiable on  $C$ , can we characterize its convexity in terms of  $\nabla f$ ?
- When  $f$  is convex on  $C$ , what can be said about its first and second differentiability?

We start with the first two questions.

### 4.1 Differentiable Convex Functions

First we assume that  $f$  is differentiable on  $C$ . Given  $x_0 \in C$ , the sentence “ $f$  is differentiable at  $x_0$ ” is meaningful only if  $f$  is at least defined in a neighborhood of  $x_0$ . Then, it is normal to assume that  $C$  is contained in an open set  $\Omega$  on which  $f$  is differentiable.

**Theorem 4.1.1** *Let  $f$  be a function differentiable on an open set  $\Omega \subset \mathbb{R}^n$ , and let  $C$  be a convex subset of  $\Omega$ . Then*

- (i)  *$f$  is convex on  $C$  if and only if*

$$f(x) \geq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle \quad \text{for all } (x_0, x) \in C \times C; \quad (4.1.1)$$

- (ii)  *$f$  is strictly convex on  $C$  if and only if strict inequality holds in (4.1.1) whenever  $x \neq x_0$ ;*

- (iii)  *$f$  is strongly convex with modulus  $c$  on  $C$  if and only if, for all  $(x_0, x) \in C \times C$ ,*

$$f(x) \geq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2}c\|x - x_0\|^2. \quad (4.1.2)$$

PROOF. [(i)] Let  $f$  be convex on  $C$ : for arbitrary  $(x_0, x) \in C \times C$  and  $\alpha \in ]0, 1[$ , we have from the definition (1.1.1) of convexity

$$f(\alpha x + (1 - \alpha)x_0) - f(x_0) \leq \alpha[f(x) - f(x_0)].$$

Divide by  $\alpha$  and let  $\alpha \downarrow 0$ : observing that  $\alpha x + (1 - \alpha)x_0 = x_0 + \alpha(x - x_0)$ , the left-hand side tends to  $\langle \nabla f(x_0), x - x_0 \rangle$  and (4.1.1) is established.

Conversely, take  $x_1$  and  $x_2$  in  $C$ ,  $\alpha \in ]0, 1[$  and define  $x_0 := \alpha x_1 + (1 - \alpha)x_2 \in C$ . By assumption,

$$f(x_i) \geq f(x_0) + \langle \nabla f(x_0), x_i - x_0 \rangle \quad \text{for } i = 1, 2 \quad (4.1.3)$$

and we obtain by convex combination

$$\alpha f(x_1) + (1 - \alpha)f(x_2) \geq f(x_0) + \langle \nabla f(x_0), \alpha x_1 + (1 - \alpha)x_2 - x_0 \rangle$$

which, after simplification, is just the relation of definition (1.1.1).

[(ii)] If  $f$  is strictly convex, we have for  $x_0 \neq x$  in  $C$  and  $\alpha \in ]0, 1[$ ,

$$f(x_0 + \alpha(x - x_0)) - f(x_0) < \alpha[f(x) - f(x_0)];$$

but  $f$  is in particular convex and we can use (i):

$$\langle \nabla f(x_0), \alpha(x - x_0) \rangle \leq f(x_0 + \alpha(x - x_0)) - f(x_0),$$

so the required strict inequality follows.

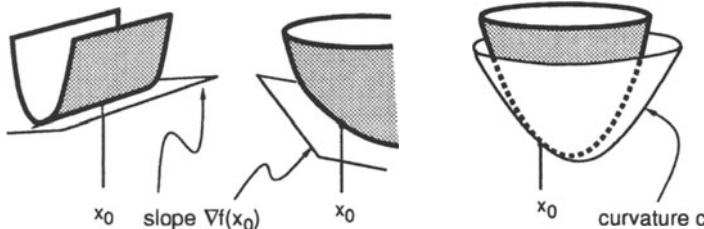
For the converse, proceed as for (i), starting from strict inequalities in (4.1.3).

[(iii)] Using Proposition 1.1.2, just apply (i) to the function  $f - \frac{1}{2}c\|\cdot\|^2$ , which is of course differentiable.  $\square$

Thus, a differentiable function is convex when its graph lies above its tangent hyperplanes: for each  $x_0$ ,  $f$  is minorized by its affine approximation  $x \mapsto f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle$  (which coincides with  $f$  at  $x_0$ ). It is strictly convex when the coincidence set is reduced to the singleton  $(x_0, f(x_0))$ . Finally,  $f$  is strongly convex when it is minorized by the quadratic convex function

$$x \mapsto f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2}c\|x - x_0\|^2,$$

whose gradient at  $x_0$  is also  $\nabla f(x_0)$ . These tangency properties are illustrated on Fig. 4.1.1.



**Fig. 4.1.1.** Affine and quadratic minorizations

**Remark 4.1.2** Inequality (4.1.1) is fundamental. In case of convexity, the remainder term  $r$  in

$$f(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + r(x_0, x)$$

must be well-behaved; for example, it is nonnegative for all  $x$  and  $x_0$ ; also,  $r(x_0, \cdot)$  is convex.  $\square$

Both  $f$ - and  $\nabla f$ -values appear in the relations dealt with in Theorem 4.1.1; we now proceed to give additional relations, involving  $\nabla f$  only. We have seen in Chap. I that a differentiable function is convex if and only if its derivative is monotone increasing (on the interval where the function is studied). Here, we need a generalization of the wording “monotone increasing” to our multidimensional situation. There are several possibilities, one is particularly well-suited to convexity:

**Definition 4.1.3** Let  $C \subset \mathbb{R}^n$  be convex. The mapping  $F : C \rightarrow \mathbb{R}^n$  is said *monotone* [resp. strictly monotone, resp. strongly monotone with modulus  $c > 0$ ] on  $C$  when, for all  $x$  and  $x'$  in  $C$ ,

$$\begin{aligned} & \langle F(x) - F(x'), x - x' \rangle \geq 0 \\ & [\text{resp. } \langle F(x) - F(x'), x - x' \rangle > 0 \quad \text{whenever } x \neq x', \\ & \quad \text{resp. } \langle F(x) - F(x'), x - x' \rangle \geq c\|x - x'\|^2 ]. \end{aligned} \quad \square$$

In the univariate case, the present monotonicity thus corresponds to  $F$  being *increasing*. When particularized to a gradient mapping  $F = \nabla f$ , our definition characterizes the convexity of the underlying potential function  $f$ :

**Theorem 4.1.4** Let  $f$  be a function differentiable on an open set  $\Omega \subset \mathbb{R}^n$ , and let  $C$  be a convex subset of  $\Omega$ . Then,  $f$  is convex [resp. strictly convex, resp. strongly convex with modulus  $c$ ] on  $C$  if and only if its gradient  $\nabla f$  is monotone [resp. strictly monotone, resp. strongly monotone with modulus  $c$ ] on  $C$ .

PROOF. We combine the “convex  $\Leftrightarrow$  monotone” and “strongly convex  $\Leftrightarrow$  strongly monotone” cases by accepting the value  $c = 0$  in the relevant relations such as (4.1.2).

Thus, let  $f$  be [strongly] convex on  $C$ : in view of Theorem 4.1.1, we can write for arbitrary  $x_0$  and  $x$  in  $C$ :

$$\begin{aligned} f(x) &\geq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2}c\|x - x_0\|^2 \\ f(x_0) &\geq f(x) + \langle \nabla f(x), x_0 - x \rangle + \frac{1}{2}c\|x_0 - x\|^2, \end{aligned}$$

and mere addition shows that  $\nabla f$  is [strongly] monotone.

Conversely, let  $(x_0, x_1)$  be a pair of elements in  $C$ . Consider the univariate function  $t \mapsto \varphi(t) := f(x_t)$ , where  $x_t := x_0 + t(x_1 - x_0)$ ; for  $t$  in an open interval containing  $[0,1]$ ,  $x_t \in \Omega$  and  $\varphi$  is well-defined and differentiable; its derivative at  $t$  is  $\varphi'(t) = \langle \nabla f(x_t), x_1 - x_0 \rangle$ . Thus, we have for all  $0 \leq t' < t \leq 1$

$$\begin{aligned} \varphi'(t) - \varphi'(t') &= \langle \nabla f(x_t) - \nabla f(x_{t'}), x_1 - x_0 \rangle \\ &= \frac{1}{t-t'} \langle \nabla f(x_t) - \nabla f(x_{t'}), x_t - x_{t'} \rangle \end{aligned} \quad (4.1.4)$$

and the monotonicity relation for  $\nabla f$  shows that  $\varphi'$  is increasing,  $\varphi$  is therefore convex (Corollary I.5.3.2).

For strong convexity, set  $t' = 0$  in (4.1.4) and use the strong monotonicity relation for  $\nabla f$ :

$$\varphi'(t) - \varphi'(0) \geq \frac{1}{t}c\|x_t - x_0\|^2 = tc\|x_1 - x_0\|^2. \quad (4.1.5)$$

Because the differentiable convex function  $\varphi$  is the integral of its derivative, we can write

$$\varphi(1) - \varphi(0) - \varphi'(0) = \int_0^1 [\varphi'(t) - \varphi'(0)] dt \geq \frac{1}{2}c\|x_1 - x_0\|^2$$

which, by definition of  $\varphi$ , is just (4.1.2) (the coefficient  $1/2$  is  $\int_0^1 t dt$ !).

The same technique proves the “strictly monotone  $\Leftrightarrow$  strictly convex” case; then, (4.1.5) becomes a strict inequality – with  $c = 0$  – and remains so after integration.  $\square$

The attention of the reader is drawn on the coefficient  $c$  – and not  $1/2 c$  – in the definition 4.1.3 of strong monotonicity. Actually, a sensible rule is: “Use  $1/2$  when dealing with a square”; here, the scalar product  $\langle \Delta F, \Delta x \rangle$  is homogeneous to a square. Alternatively, remember in Proposition 1.1.2 that the gradient of  $1/2 c\|\cdot\|^2$  at  $x$  is  $cx$ .

We mention the following example: let

$$f(x) := \frac{1}{2}\langle Ax, x \rangle + \langle b, x \rangle$$

be a quadratic convex function ( $A$  is symmetric), and let  $\lambda_n \geq 0$  be its smallest eigenvalue. Observe that  $\nabla f(x) = Ax + b$  and that

$$\langle Ax - Ax', x - x' \rangle = \langle A(x - x'), x - x' \rangle \geq \lambda_n\|x - x'\|^2.$$

Thus  $\nabla f$  is monotone [strongly with modulus  $\lambda_n$ ]. The [strong] convexity of  $f$ , in the sense of (1.1.2), has been already alluded to in §1.3(d); but (4.1.2) is easier to establish here: simply write

$$\begin{aligned} f(x) - f(x_0) - \langle \nabla f(x_0), x - x_0 \rangle &= \frac{1}{2}\langle Ax, x \rangle - \frac{1}{2}\langle Ax_0, x_0 \rangle - \langle Ax_0, x - x_0 \rangle \\ &= \frac{1}{2}\langle A(x - x_0), x - x_0 \rangle \geq \frac{1}{2}\lambda_n\|x - x_0\|^2. \end{aligned}$$

Note that for this particular class of convex functions, strong and strict convexity are equivalent to each other, and to the positive definiteness of  $A$ .

**Remark 4.1.5** Do not infer from Theorem 4.1.4 the statement “a monotone mapping is the gradient of a convex function”, which is wrong. To be so, the mapping in question must first be a gradient, an issue that we do not study here. We just mention the following property: if  $\Omega$  is convex and  $F : \Omega \rightarrow \mathbb{R}^n$  is differentiable, then  $F$  is a gradient if and only if its Jacobian operator is symmetric (in 2 or 3 dimensions,  $\text{curl } F \equiv 0$ ).  $\square$

**Example 4.1.6** Let  $C \subset \mathbb{R}^n$  be nonempty closed convex. We have already seen in Example 2.1.4 that the function

$$\mathbb{R}^n \ni x \mapsto \varphi_C(x) := \frac{1}{2}[\|x\|^2 - d_C^2(x)]$$

is convex and finite everywhere. It would be so for arbitrary  $C$ , but the convexity of  $C$  here implies the differentiability of  $\varphi_C$ , with gradient  $\nabla \varphi_C = p_C$  (the projection operator on  $C$ ). To differentiate the only delicate term  $d_C^2$ , consider

$$\Delta := d_C^2(x + h) - d_C^2(x).$$

Because  $d_C^2(x) \leq \|x - p_C(x + h)\|^2$ , we have

$$\Delta \geq \|x + h - p_C(x + h)\|^2 - \|x - p_C(x + h)\|^2 = \|h\|^2 + 2\langle h, x - p_C(x + h) \rangle.$$

Inverting the role of  $x$  and  $x + h$ , we obtain likewise

$$\Delta \leq \|x + h - p_C(x)\|^2 - \|x - p_C(x)\|^2 = \|h\|^2 + 2\langle h, x - p_C(x) \rangle.$$

Now remember from (III.3.1.6) that  $p_C$  is nonexpansive, hence

$$\Delta = 2\langle x - p_C(x), h \rangle + o(\|h\|),$$

and this gives the announced result. Incidentally, take  $x \notin C$ , hence  $d_C(x) > 0$ ; by standard differential calculus,

$$\nabla d_C(x) = \nabla \sqrt{d_C^2}(x) = \frac{x - p_C(x)}{\|x - p_C(x)\|}.$$

Now we ask the question: is  $\varphi_C$  strictly or strongly convex on some (nontrivial) convex set? Because of Proposition III.3.1.3, we have for all  $x$  and  $x'$  in  $C$ :

$$\begin{aligned} [\langle \nabla \varphi_C(x) - \nabla \varphi_C(x'), x - x' \rangle] &= \langle p_C(x) - p_C(x'), x - x' \rangle \\ &\geq \|p_C(x) - p_C(x')\|^2 = \|x - x'\|^2, \end{aligned} \quad (4.1.6)$$

so  $\nabla \varphi_C$  [resp.  $\varphi_C$ ] is strongly monotone [resp. strongly convex] with modulus 1 on  $C$  – but we knew it already, since  $\varphi_C = \frac{1}{2}\|\cdot\|^2$  there.

On the other hand,  $\varphi_C$  cannot be strongly convex outside  $C$ : take  $p \in \text{bd } C$  and two different points  $x, x'$  in the normal cone  $N_C(p)$ ; then  $p_C(x) = p_C(x') = p$  and the left-hand side of (4.1.6) is zero. In other words,  $\varphi_C$  is affine on  $\{p\} + N_C(p)$ . This geometrical property is illustrated by Fig. 4.1.2: apply the triangular relation

$$\|p\|^2 + \|x - p\|^2 = \|x\|^2 - 2\|p\| \|x - p\| \cos \theta$$

to observe that

$$\varphi_C(x) = \frac{1}{2}(\|x\|^2 - \|x - p\|^2) = \frac{1}{2}\|p\|^2 + \|p\| \|x - p\| \cos \theta$$

is affine with respect to the single variable  $\|x - p\|$  when  $p$  and the angle  $\theta$  are fixed.  $\square$

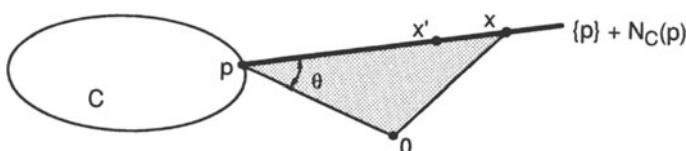


Fig. 4.1.2. Difference of squared distances

## 4.2 Nondifferentiable Convex Functions

A convex function need not be differentiable over the whole interior of its domain; nevertheless, it is so at “many points” in this set. Before making this sentence mathematically precise, we note the following nice property of convex functions.

**Proposition 4.2.1** *For  $f \in \text{Conv } \mathbb{R}^n$  and  $x \in \text{int dom } f$ , the three statements below are equivalent:*

(i) *The function*

$$\mathbb{R}^n \ni d \mapsto \lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t} \text{ is linear in } d;$$

(ii) *for some basis of  $\mathbb{R}^n$  in which  $x = (\xi^1, \dots, \xi^n)$ , the partial derivatives  $\frac{\partial f}{\partial \xi^i}(x)$  exist at  $x$ , for  $i = 1, \dots, n$ ;*

(iii)  *$f$  is differentiable at  $x$ .*

PROOF. First of all, remember from Theorem I.4.1.1 that the one-dimensional function  $t \mapsto f(x + td)$  has half-derivatives at 0: the limits considered in (i) exist for all  $d$ . We will denote by  $\{b_1, \dots, b_n\}$  the basis postulated in (ii), so that  $x = \sum_{i=1}^n \xi^i b_i$ .

Denote by  $d \mapsto \ell(d)$  the function defined in (i); taking  $d = \pm b_i$ , realize that, when (i) holds,

$$\lim_{\tau \uparrow 0} \frac{f(x + \tau b_i) - f(x)}{-\tau} = \ell(-b_i) = -\ell(b_i) = -\lim_{t \downarrow 0} \frac{f(x + tb_i) - f(x)}{t}.$$

This means that the two half-derivatives at  $t = 0$  of the function  $t \mapsto f(x + tb_i)$  coincide: the partial derivative of  $f$  at  $x$  along  $b_i$  exists, (ii) holds. That (iii) implies (i) is clear: when  $f$  is differentiable at  $x$ ,

$$\lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t} = \langle \nabla f(x), d \rangle.$$

We do not really complete the proof here, because everything follows in a straightforward way from subsequent chapters. More precisely, [(ii)  $\Rightarrow$  (i)] is Proposition V.1.1.6, which states that the function  $\ell$  is linear on the space generated by the  $b_i$ 's, whenever it is linear along each  $b_i$ . Finally [(i)  $\Rightarrow$  (iii)] results from Lemma VI.2.1.1 and the proof goes as follows. One of the possible definitions of (iii) is:

$$\lim_{t \downarrow 0, d' \rightarrow d} \frac{f(x + td') - f(x)}{t} \text{ is linear in } d.$$

Because  $f$  is locally Lipschitzian, the above limit exists whenever it exists for fixed  $d' = d$  – i.e. the expression in (i).  $\square$

The function defined in (i), called the *directional derivative* of  $f$  at  $x$  in the direction  $d$ , is denoted by

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

and will be encountered many times in the sequel.

**Remark 4.2.2** The above result reveals three interesting properties enjoyed by convex functions:

- First, consider the restriction  $\varphi_d(t) := f(x + td)$  of  $f$  along a line  $x + \mathbb{R}d$ . As soon as there are  $n$  independent directions, say  $d_1, \dots, d_n$ , such that each  $\varphi_{d_i}$  has a derivative at  $t = 0$ , then the same property holds for all the other possible directions  $d \in \mathbb{R}^n$ .
- Second, this “radial” differentiability property of  $\varphi_d$  for all  $d$  (or for many enough  $d$ ) suffices to guarantee the “global” (i.e. Fréchet) differentiability of  $f$  at  $x$ ; a property which does not hold in general. It depends crucially on the Lipschitz property of  $f$ .
- One can also show that, if  $f$  is convex and differentiable in a neighborhood of  $x$ , then  $\nabla f$  is continuous at  $x$ . Hence, if  $\Omega$  is an open convex set, the following equivalence holds true for the convex  $f$ :

$$f \text{ differentiable on } \Omega \iff f \in C^1(\Omega).$$

This rather surprising property will be confirmed in §VI.6.2. □

The largest set on which a function can be differentiable is the interior of its domain. We are now in a position to show that a convex function is differentiable almost everywhere on that set.

**Theorem 4.2.3** *Let  $f \in \text{Conv } \mathbb{R}^n$ . The subset of  $\text{int dom } f$  where  $f$  fails to be differentiable is of zero (Lebesgue) measure.*

PROOF. Since  $\text{int dom } f$  is the union, for  $k = 1, 2, \dots$ , of the open sets

$$\Omega_k := \{x \in \text{int dom } f : \|x\| < k\},$$

it suffices to prove that each set

$$E := \{x \in \Omega_k : f \text{ is not differentiable at } x\}$$

is of measure zero. In view of Proposition 4.2.1,  $E$  is also the set where some partial derivative does not exist. In other words

$$E = E_1 \cup \dots \cup E_n;$$

here  $E_i$  is the subset of  $\Omega_k$  where the partial derivative of  $f$  along  $b_i$  does not exist ( $\{b_1, \dots, b_n\}$  being some basis of  $\mathbb{R}^n$ ). Using the property (I.4.1.3) of increasing slopes,

$$E_i := \{x \in \Omega_k : f'(x, b_i) > -f'(x, -b_i)\}.$$

Each  $E_i$  is measurable (the functions  $f'(\cdot, d)$  are measurable as pointwise limits of measurable functions), so we will be done if we prove that each  $E_i$  is of measure zero.

Let us establish  $\text{meas } E_1 = 0$ . First,  $E_1 \subset \Omega_k$  is bounded, so the characteristic function  $\chi_1$  (1 on  $E_1$ , 0 elsewhere) is integrable. According to Fubini's Theorem (§A.6.2), we therefore write the measure of  $E_1$  as

$$\int \chi_1(\xi^1, \dots, \xi^n) d\xi^1 \dots d\xi^n = \int \dots \int \left[ \int \chi_1(\xi^1, \dots, \xi^n) d\xi^1 \right] d\xi^2 \dots d\xi^n,$$

$(\xi^1, \dots, \xi^n)$  being the coordinates of a point  $x$  along the given basis. The one-dimensional convex function  $\xi^1 \mapsto f(\xi^1, \xi^2, \dots, \xi^n)$  has a derivative except possibly at countably many points of  $E_1$  (Theorem I.4.2.1), and this implies

$$\int \chi_1(\xi^1, \dots, \xi^n) d\xi^1 = 0 \quad \text{for all } \xi^2, \dots, \xi^n. \quad \square$$

It is worth mentioning that this property follows from a more general result, due to H. Rademacher (1919): a function which is locally Lipschitzian on an open set  $\Omega$ , for example a convex function (Theorem 3.1.2), is differentiable almost everywhere in  $\Omega$ . Our direct proof above cannot be extended to that case, though: we explicitly used the equivalence (ii)  $\Leftrightarrow$  (iii) of Proposition 4.2.1, for which convexity is essential.

### 4.3 Second-Order Differentiation

We have seen in Chap. I that the most useful criterion to recognize a convex function uses the second derivative: a function  $\varphi$  which is twice differentiable on an interval  $I$  is convex on  $I$  if and only if  $\varphi''$  is nonnegative on  $I$ . In our present framework, the best idea is to reduce the question to the one-dimensional case: a function is convex if and only if its restrictions to the segments  $[x, x']$  are also convex. These segments can in turn be parametrized via an origin  $x$  and a direction  $d$ : convexity of  $f$  amounts to the convexity of  $t \mapsto f(x + td)$ . Then, it suffices to apply calculus rules to compute the second derivative of this last function. Our first result mimics Theorem 4.1.4.

**Theorem 4.3.1** *Let  $f$  be twice differentiable on an open convex set  $\Omega \subset \mathbb{R}^n$ . Then*

- (i)  *$f$  is convex on  $\Omega$  if and only if  $\nabla^2 f(x_0)$  is positive semi-definite for all  $x_0 \in \Omega$ ;*
- (ii) *if  $\nabla^2 f(x_0)$  is positive definite for all  $x_0 \in \Omega$ , then  $f$  is strictly convex on  $\Omega$ ;*
- (iii)  *$f$  is strongly convex with modulus  $c$  on  $\Omega$  if and only if the smallest eigenvalue of  $\nabla^2 f(\cdot)$  is minorized by  $c$  on  $\Omega$ : for all  $x_0 \in \Omega$  and all  $d \in \mathbb{R}^n$ ,*

$$\langle \nabla^2 f(x_0)d, d \rangle \geq c\|d\|^2.$$

PROOF. For given  $x_0 \in \Omega$ ,  $d \in \mathbb{R}^n$  and  $t \in \mathbb{R}$  such that  $x_0 + td \in \Omega$ , we will set

$$x_t := x_0 + td \quad \text{and} \quad \varphi(t) := f(x_t) = f(x_0 + td),$$

so that  $\varphi''(t) = \langle \nabla^2 f(x_t)d, d \rangle$ .

[(i)] Assume  $f$  is convex on  $\Omega$ ; let  $(x_0, d)$  be arbitrary in  $\Omega \times \mathbb{R}^n$ , with  $d \neq 0$ :  $\varphi$  is then convex on the open interval  $I := \{t \in \mathbb{R} : x_0 + td \in \Omega\}$ . It follows

$$0 \leq \varphi''(t) = \langle \nabla^2 f(x_t)d, d \rangle \quad \text{for all } t \in I \ni 0$$

and  $\nabla^2 f(x_0)$  is positive semi-definite.

Conversely, take an arbitrary  $[x_0, x_1] \subset \Omega$ , set  $d := x_1 - x_0$  and assume  $\nabla^2 f(x_t)$  positive semi-definite, i.e.  $\varphi''(t) \geq 0$ , for  $t \in [0, 1]$ . Then Theorem I.5.3.3 tells us that  $\varphi$  is convex on  $[0, 1]$ , i.e.  $f$  is convex on  $[x_0, x_1]$ . The result follows since  $x_0$  and  $x_1$  were arbitrary in  $\Omega$ .

[*(ii)*] To establish the strict convexity of  $f$  on  $\Omega$ , we prove that  $\nabla f$  is strictly monotone on  $\Omega$ : Theorem 4.1.4 will apply. As above, take an arbitrary  $[x_0, x_1] \subset \Omega$ ,  $x_1 \neq x_0$ ,  $d := x_1 - x_0$ , and apply the mean-value theorem to the function  $\varphi'$ , differentiable on  $[0, 1]$ : for some  $\tau \in ]0, 1[$ ,

$$\varphi'(1) - \varphi'(0) = \varphi''(\tau) = \langle \nabla^2 f(x_\tau)d, d \rangle > 0$$

and the result follows since

$$\varphi'(1) - \varphi'(0) = \langle \nabla f(x_1) - \nabla f(x_0), x_1 - x_0 \rangle.$$

[*(iii)*] Using Proposition 1.1.2, apply (i) to the function  $f - \frac{1}{2}c\|\cdot\|^2$ , whose Hessian operator is  $\nabla^2 f - cI_n$  and has the eigenvalues  $\lambda - c$ , with  $\lambda$  describing the eigenvalues of  $\nabla^2 f$ .  $\square$

Some differences have appeared with respect to §4.1:

- The sufficiency condition in (ii) is not necessary, even for univariate functions: think of  $f(x) = \frac{1}{4}x^4$ .
- Theorem 4.1.1 stated that the affine (first-order) approximation of  $f$  around  $x_0$  was actually a global minorization – more or less “comfortable”. Here, we cannot say that the quadratic approximation (of  $f$  around  $x_0$ )

$$x \mapsto f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + \frac{1}{2}\langle \nabla^2 f(x_0)(x - x_0), x - x_0 \rangle$$

minorizes  $f$ : think of  $f(x) = \frac{1}{2}x^2 - \frac{1}{4}x^4$ , which is convex for  $|x|^2 \leq \frac{1}{3}$ .

- The present statements do not characterize convexity on a convex subset  $C \subset \Omega$ :  $C$  must be *open*. The reason is that §4.1 was dealing with the image (through  $f$  or  $\nabla f$ ) of pairs of points in  $C$  ( $x_0$  and  $x$ , or  $x$  and  $x'$ ). Here,  $\nabla^2 f$  looks at  $f$  in the neighborhood of a single point, say  $x_0$ . Thus, a statement like

$$f \text{ is convex on } C \subset \Omega \iff \nabla^2 f(\cdot) \text{ is positive semi-definite on } C$$

may be wrong if  $C$  is not open:  $f(\xi, \eta) := \xi^2 - \eta^2$  is convex on  $C = \mathbb{R} \times \{0\}$  but its Hessian is nowhere positive semi-definite.

**Remark 4.3.2** Despite the last comment above, the convexity criterion using second derivatives is still the most powerful, even if positive (semi-)definiteness is not always easy to check. To recognize a convex function on a non-open set  $C$ , the best chance is therefore to use the Hessian operator on  $\Omega = \text{int } C$ , hopefully nonempty, and then to try and conclude by passing to the limit: the property  $C \subset \text{cl}(\text{int } C) = \text{cl } C$  is useful for that.  $\square$

**Example 4.3.3** To illustrate Theorem 4.3.1, consider the function

$$\Omega := \left\{ x = (\xi^1, \dots, \xi^n) : \xi^i > 0 \text{ for } i = 1, \dots, n \right\},$$

$$f : \Omega \ni x \mapsto f(x) := -(\xi^1 \xi^2 \cdots \xi^n)^{1/n}.$$

Direct computations give its second derivatives, i.e. its Hessian operator associated with the dot-product of  $\mathbb{R}^n$ :

$$\frac{\partial^2 f}{\partial \xi^i \partial \xi^j}(x) = \frac{f(x)}{n^2 \xi^i \xi^j} (1 - n \delta_{ij})$$

where  $\delta_{ij}$  is Kronecker's symbol. We obtain, with  $d = (d^1, \dots, d^n) \in \mathbb{R}^n$ :

$$[\nabla^2 f(x)d]^\top d = \frac{f(x)}{n^2} \left[ \left( \sum_{i=1}^n \frac{d^i}{\xi^i} \right)^2 - n \sum_{i=1}^n \left( \frac{d^i}{\xi^i} \right)^2 \right].$$

The  $\ell_1$ - and  $\ell_2$ -norms are related on  $\mathbb{R}^n$  by the inequality  $\|\cdot\|_1 \leq \sqrt{n} \|\cdot\|_2$  (take a vector of the type  $(\pm 1, \dots, \pm 1)$  and use the Cauchy-Schwarz inequality); because  $f$  is negative on  $\Omega$ , the above expression is therefore nonnegative:  $f$  is convex. Observe in passing that we obtain an equality if  $d$  and  $x$  are collinear: our  $f$  is positively homogeneous.

Observe here that  $f$  can be extended to  $\text{cl } \Omega$  by posing  $f(x) = 0$  if some  $\xi^i$  is zero. Convexity is preserved, and this illustrates Remark 4.3.2.  $\square$

Having thus established a parallel with §4.1, we now consider the *existence* of second derivatives. In one dimension, the first derivative is monotone and, as such, it in turn has a derivative almost everywhere (Theorem I.5.1.3). In several dimensions, monotonicity becomes that of Definition 4.1.3; the differentiability of such operators involves much more sophisticated concepts from analysis. We just mention without proof the main result:

**Theorem 4.3.4 (A.D. Alexandrov)** *Let  $f \in \text{Conv } \mathbb{R}^n$ . For all  $x \in \text{int dom } f$  except in a set of zero (Lebesgue) measure,  $f$  is differentiable at  $x$  and there exists a symmetric positive semi-definite operator  $D^2 f(x)$  such that, for  $h \in \mathbb{R}^n$*

$$f(x + h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} \langle D^2 f(x)h, h \rangle + o(\|h\|^2). \quad \square$$

The operator  $D^2 f(x)$  can hardly be called the “second derivative” of  $f$  at  $x$  because its existence does not even imply the existence of  $\nabla f$  in a neighborhood of  $x$ . One should rather say that it gives a *second-order approximation* of  $f$  around  $x$ .

**Remark 4.3.5** We also mention that much can be said concerning the set  $E_1$  where  $\nabla f$  fails to exist; but the set  $E_2$  where  $D^2 f$  fails to exist seems much more mysterious. Despite the analogy between Theorems 4.2.3 and 4.3.4, there is a drastic difference between first- and second-order approximations of a convex function.

Also, the interesting properties mentioned in Remark 4.2.2 do not transfer to second order:  $f$  may be twice differentiable on  $\Omega$  without being twice continuously differentiable on  $\Omega$ ; the (first- and) second-order partial derivatives of  $f$  may exist at  $x$  while  $f$  is not twice differentiable at  $x$ ; and so on.  $\square$

**Remark 4.3.6 (The Case of Flat Domains)** In all the present Section 4,  $\text{dom } f$  was implicitly assumed full-dimensional, in order to have a nonempty  $\text{int dom } f$ . When such is not the case, some kind of differentiation can still be performed. In fact, exploit Remark 2.1.7 and make a change of variable:

$$y \mapsto f_0(y) := f(x_0 + y),$$

where  $x_0$  is fixed in  $\text{dom } f$ ,  $y$  varies in  $V$ ,  $V \subset \mathbb{R}^n$  being the subspace parallel to  $\text{aff dom } f$ . Now,  $f_0 \in \text{Conv } V$  and  $\text{dom } f_0$  is full-dimensional in  $V$ . Equipping  $V$  with the induced scalar product  $\langle \cdot, \cdot \rangle$  and the induced Lebesgue measure, the main results above can be reproduced. More precisely: almost everywhere in  $\text{int dom } f_0$ , i.e. for almost all  $x_0 + y \in \text{ri dom } f$ ,

- there is a vector  $s \in V$  (the gradient of  $f_0$  at  $y$ ) which gives the first-order approximation of  $f$  around  $x_0$

$$\forall h \in V, \quad f(x_0 + y + h) = f(x_0 + y) + \langle s, h \rangle + o(\|h\|)$$

(the remainder term  $o(\|h\|)$  being nonnegative); this  $s$  could be called the “relative gradient” of  $f$  at  $x := x_0 + y$ ; it exists at  $x_0 + y$  if and only if the function  $t \mapsto f(x_0 + y + td)$  has a derivative at  $t = 0$  for all  $d \in V$ ;

- there is a linear operator  $D : V \rightarrow V$  which is symmetric positive semi-definite: for all  $h$  and  $k$  in  $V$ ,

$$Dh \in V \quad \text{and} \quad \langle Dh, k \rangle = \langle h, Dk \rangle,$$

and which gives the second-order approximation

$$\forall h \in V, \quad f(x_0 + y + h) = f(x_0 + y) + \langle s, h \rangle + \langle Dh, h \rangle + o(\|h\|^2). \quad \square$$

## V. Sublinearity and Support Functions

**Prerequisites.** Basic definitions, properties and operations of convex sets (Chap. III) and convex functions (Chap. IV).

**Introduction.** In classical real analysis, the simplest functions are *linear*. In convex analysis, the next simplest convex functions (apart from the affine functions, widely used in §IV.1.2), are so-called *sublinear*. There are several motivations for their study; we give three of them, which are of particular importance in the context of convex analysis and optimization.

(i) *A suitable generalization of linearity.* A linear function  $\ell$  from  $\mathbb{R}^n$  to  $\mathbb{R}$ , or a linear form on  $\mathbb{R}^n$ , is primarily defined as a function satisfying for all  $(x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n$  and  $(t_1, t_2) \in \mathbb{R} \times \mathbb{R}$ :

$$\ell(t_1 x_1 + t_2 x_2) = t_1 \ell(x_1) + t_2 \ell(x_2). \quad (0.1)$$

A corresponding definition for a sublinear function  $\sigma$  from  $\mathbb{R}^n$  into  $\mathbb{R}$  is: for all  $(x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n$  and  $(t_1, t_2) \in \mathbb{R}^+ \times \mathbb{R}^+$ ,

$$\sigma(t_1 x_1 + t_2 x_2) \leq t_1 \sigma(x_1) + t_2 \sigma(x_2). \quad (0.2)$$

A first observation is that requiring an inequality in (0.2), rather than an equality, allows infinite values for  $\sigma$  without destroying the essence of the concept of sublinearity. Of course, (0.2) is less stringent than (0.1), but more stringent than the definition of a convex function: the inequality must hold in (0.2) even if  $t_1 + t_2 \neq 1$ . This confirms that sublinear functions, which generalize linear functions, are particular instances of convex functions.

**Remark 0.1** Note that (0.1) and (0.2) can be made more similar by restricting  $t_1$  and  $t_2$  to be positive in (0.1) – this leaves unchanged the definition of a linear function.

The prefix “sub” comes from the inequality-sign “ $\leq$ ” in (0.2). It also suggests that sublinearity is less demanding than linearity, but this is a big piece of luck. In fact, draw the graph of a convex and of a concave function and ask a non-mathematician: “which is convex?”. He will probably give the wrong answer. Yet, if convex functions were defined the other way round, (0.2) should have the “ $\geq$ ” sign. The associated concept would be superlinearity, an unfortunate wording which suggests “more than linear”.  $\square$

In a word, sublinear functions are reasonable candidates for “simplest non-trivial convex functions”. Whether they are interesting candidates will be seen in (ii) and (iii). Here, let us just mention that their epigraph is a convex cone, the next simplest convex epigraph after half-spaces.

(ii) *Tangential approximation of convex functions.* To say that a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is differentiable at  $x$  is to say that there is a linear function  $\ell_x$  which approximates  $f(x + h) - f(x)$  to first order, i.e.

$$f(x + h) - f(x) = \ell_x(h) + o(\|h\|).$$

This fixes the rate of change of  $f$  when  $x$  is moved along a line  $d$ : with  $\varepsilon(t) \rightarrow 0$  if  $t \rightarrow 0$ ,

$$\frac{f(x + td) - f(x)}{t} = \ell_x(d) + \varepsilon(t) \quad \text{for all } t \neq 0.$$

Geometrically, the graph of  $f$  has a tangent hyperplane at  $(x, f(x)) \in \mathbb{R}^n \times \mathbb{R}$ ; and this hyperplane is the graph  $\text{gr } \ell_x$  of the affine function  $h \mapsto f(x) + \ell_x(h)$ .

When  $f$  is merely convex, its graph may have no tangent hyperplane at a given  $(x, f(x))$ . Nevertheless, under reasonable assumptions,  $f(x + h) - f(x)$  can still be approximated to first order by a function which is sublinear: there exists a sublinear function  $h \mapsto \sigma_x(h)$  such that

$$f(x + h) - f(x) = \sigma_x(h) + o(\|h\|).$$

This will be seen in Chap. VI.

Geometrically,  $\text{gr } \sigma_x$  is no longer a hyperplane but rather a cone, which is therefore tangent to  $\text{gr } f$  (the word “tangent” should be understood here in its intuitive meaning of a tangent surface, as opposed to tangent cones of Chap. III; neither  $\text{gr } \sigma_x$  nor  $\text{gr } f$  are convex). Thus, one can say that differentiable functions are “tangentially linear”, while convex functions are “tangentially sublinear”. See Fig. 0.1, which displays the graph of a differentiable and of a convex function. The graph of  $\ell_x$  is the thick line  $L$ , while the graph of  $\sigma_x$  is made up of the two thick half-lines  $S_1$  and  $S_2$ .

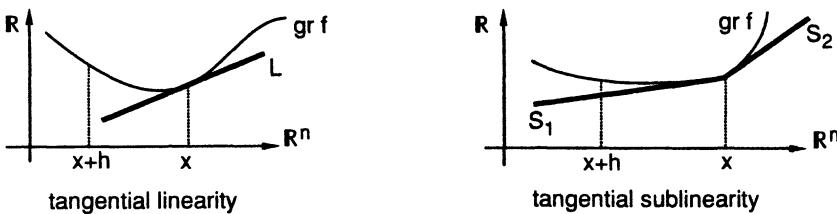


Fig. 0.1. Two concepts of tangency

(iii) *Nice correspondence with nonempty closed convex sets.* In the Euclidean space  $(\mathbb{R}^n, \langle \cdot, \cdot \rangle)$ , a linear form  $\ell$  can be represented by a vector: there is a unique  $s \in \mathbb{R}^n$  such that

$$\ell(x) = \langle s, x \rangle \quad \text{for all } x \in \mathbb{R}^n. \quad (0.3)$$

The definition (0.3) of a linear function is more geometric than (0.1), and just as accurate. A large part of the present chapter will be devoted to generalizing the above representation theorem to sublinear functions.

First observe that, given a nonempty set  $S \subset \mathbb{R}^n$ , the function  $\sigma_S : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  defined by

$$\sigma_S(x) := \sup \{\langle s, x \rangle : s \in S\} \quad (0.4)$$

is sublinear. It is called the *support function* of  $S$ , already encountered in Sects III.4.1 and IV.1.3(a). When  $S$  is bounded, its support function is finite everywhere; otherwise,  $\sigma_S$  can take on the value  $+\infty$  but it remains lower semi-continuous. Furthermore, it is easy to check that  $\sigma_S$  is also the support function of the closure of  $S$ , and even of the closed convex hull of  $S$ . It is therefore logical to consider support functions of nonempty closed convex sets only.

Now, a key result is that the mapping  $S \mapsto \sigma_S$  is then bijective: a lower semi-continuous (i.e. closed) sublinear function is the support function of a *uniquely determined* nonempty closed convex set. Thus, (0.4) establishes the announced representation, just as (0.3) does in the linear case. Note that the linear case is covered: it corresponds to  $S$  being a singleton  $\{s\}$  in (0.4).

This correspondence between nonempty closed convex sets of  $\mathbb{R}^n$  and closed sublinear functions allows fruitful and enlightening geometric interpretations when studying these functions. Vice versa, it provides powerful analytical tools for the study of these sets. In particular, when closed convex sets are combined (intersected, added, etc.) to form new convex sets, we will show how their support functions are correspondingly combined: the mapping (0.4) is an *isomorphism*, with respect to a number of structures.

## 1 Sublinear Functions

### 1.1 Definitions and First Properties

**Definition 1.1.1** A function  $\sigma : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  is said to be *sublinear* if it is convex and positively homogeneous (of degree 1):  $\sigma \in \text{Conv } \mathbb{R}^n$  and

$$\sigma(tx) = t\sigma(x) \quad \text{for all } x \in \mathbb{R}^n \text{ and } t > 0. \quad (1.1.1)$$

□

**Remark 1.1.2** Inequality in (1.1.1) would be enough to define positive homogeneity: a function  $\sigma$  is positively homogeneous if and only if it satisfies

$$\sigma(tx) \leq t\sigma(x) \quad \text{for all } x \in \mathbb{R}^n \text{ and } t > 0. \quad (1.1.2)$$

In fact, (1.1.2) implies ( $tx \in \mathbb{R}^n$  and  $t^{-1} > 0!$ )

$$\sigma(x) = \sigma(t^{-1}tx) \leq t^{-1}\sigma(tx)$$

which, together with (1.1.2), shows that  $\sigma$  is positively homogeneous. □

We deduce from (1.1.1) that  $\sigma(0) = t\sigma(0)$  for all  $t > 0$ . This leaves only two possible values for  $\sigma(0)$ : 0 and  $+\infty$ . However, most of the sublinear functions to be encountered in the sequel do satisfy  $\sigma(0) = 0$ . According to our Definition IV.1.1.3 of convex functions,  $\sigma$  should be finite somewhere; otherwise  $\text{dom } \sigma$  would be empty. Now, if  $\sigma(x) < +\infty$ , (1.1.1) shows that  $\sigma(tx) < +\infty$  for all  $t > 0$ . In other words,  $\text{dom } \sigma$  is a cone, convex because  $\sigma$  is itself convex. Note that, being convex,  $\sigma$  is continuous relatively to  $\text{ri dom } \sigma$ , but discontinuities may occur on the boundary-rays of  $\text{dom } \sigma$ , including at 0.

The following result is a geometrical characterization of sublinear functions.

**Proposition 1.1.3** *A function  $\sigma : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  is sublinear if and only if its epigraph  $\text{epi } \sigma$  is a nonempty convex cone in  $\mathbb{R}^n \times \mathbb{R}$ .*

PROOF. We know that  $\sigma$  is a convex function if and only if  $\text{epi } \sigma$  is a nonempty convex set in  $\mathbb{R}^n \times \mathbb{R}$  (Proposition IV.1.1.6). Therefore, we just have to prove the equivalence between positive homogeneity and  $\text{epi } \sigma$  being a cone.

Let  $\sigma$  be positively homogeneous. For  $(x, r) \in \text{epi } \sigma$ , the relation  $\sigma(x) \leq r$  gives

$$\sigma(tx) = t\sigma(x) \leq tr \quad \text{for all } t > 0,$$

so  $\text{epi } \sigma$  is a cone. Conversely, if  $\text{epi } \sigma$  is a cone in  $\mathbb{R}^n \times \mathbb{R}$ , the property  $(x, \sigma(x)) \in \text{epi } \sigma$  implies  $(tx, t\sigma(x)) \in \text{epi } \sigma$ , i.e.

$$\sigma(tx) \leq t\sigma(x) \quad \text{for all } t > 0.$$

From Remark 1.1.2, this is just positive homogeneity. □

Another important concept in analysis is *subadditivity*: a function  $\sigma$  is subadditive when it satisfies

$$\sigma(x_1 + x_2) \leq \sigma(x_1) + \sigma(x_2) \quad \text{for all } (x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n. \quad (1.1.3)$$

Here again, the inequality is understood in  $\mathbb{R} \cup \{+\infty\}$ . Together with positive homogeneity, the above axiom gives another characterization (analytical, rather than geometrical) of sublinear functions.

**Proposition 1.1.4** *A function  $\sigma : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ , not identically equal to  $+\infty$ , is sublinear if and only if one of the following two properties holds:*

$$\sigma(t_1 x_1 + t_2 x_2) \leq t_1 \sigma(x_1) + t_2 \sigma(x_2) \quad \text{for all } x_1, x_2 \in \mathbb{R}^n \text{ and } t_1, t_2 > 0, \quad (1.1.4)$$

or

$$\sigma \text{ is positively homogeneous and subadditive}. \quad (1.1.5)$$

PROOF. [*sublinearity  $\Rightarrow$  (1.1.4)*] For  $x_1, x_2 \in \mathbb{R}^n$  and  $t_1, t_2 > 0$ , set  $t := t_1 + t_2 > 0$ ; we have

$$\begin{aligned} \sigma(t_1 x_1 + t_2 x_2) &= \sigma\left(t\left[\frac{t_1}{t}x_1 + \frac{t_2}{t}x_2\right]\right) \\ &= t\sigma\left(\frac{t_1}{t}x_1 + \frac{t_2}{t}x_2\right) \quad [\text{positive homogeneity}] \\ &\leq t\left[\frac{t_1}{t}\sigma(x_1) + \frac{t_2}{t}\sigma(x_2)\right], \quad [\text{convexity}] \end{aligned}$$

and (1.1.4) is proved.

$[(1.1.4) \Rightarrow (1.1.5)]$  A function satisfying (1.1.4) is obviously subadditive (take  $t_1 = t_2 = 1$ ) and satisfies (take  $x_1 = x_2 = x$ ,  $t_1 = t_2 = 1/2t$ )

$$\sigma(tx) \leq t\sigma(x) \quad \text{for all } x \in \mathbb{R}^n \text{ and } t > 0,$$

which is just positive homogeneity because of Remark 1.1.2.

$[(1.1.5) \Rightarrow \text{sublinearity}]$  Take  $t_1, t_2 > 0$  with  $t_1 + t_2 = 1$  and apply successively subadditivity and positive homogeneity:

$$\sigma(t_1x_1 + t_2x_2) \leq \sigma(t_1x_1) + \sigma(t_2x_2) = t_1\sigma(x_1) + t_2\sigma(x_2),$$

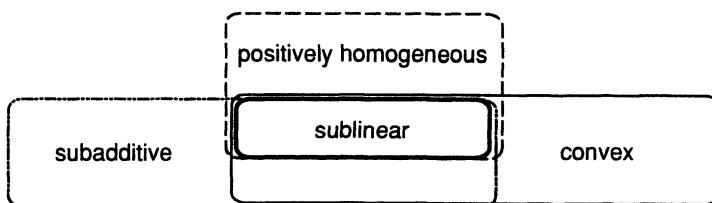
hence  $\sigma$  is convex.  $\square$

**Corollary 1.1.5** *If  $\sigma$  is sublinear, then*

$$\sigma(x) + \sigma(-x) \geq 0 \quad \text{for all } x \in \mathbb{R}^n. \quad (1.1.6)$$

PROOF. Take  $x_2 = -x_1$  in (1.1.3) and remember that  $\sigma(0) \geq 0$ .  $\square$

It is worth mentioning that, to become sublinear, a positively homogeneous function just needs to be subadditive as well (rather than convex, as suggested by Definition 1.1.1); then, of course, it becomes convex at the same time. Figure 1.1.1 summarizes the connections between the classes of functions given so far. Note for completeness that a convex and subadditive function need not be sublinear: think of  $f(x) \equiv 1$ .



**Fig. 1.1.1.** Various classes of functions

Similarly, one can ask when a sublinear function becomes linear. For a linear function, (1.1.6) holds as an equality, and the next result implies that this is exactly what makes the difference.

**Proposition 1.1.6** *Let  $\sigma$  be sublinear and suppose that there exist  $x_1, \dots, x_m$  in  $\text{dom } \sigma$  such that*

$$\sigma(x_j) + \sigma(-x_j) = 0 \quad \text{for } j = 1, \dots, m. \quad (1.1.7)$$

*Then  $\sigma$  is linear on the subspace spanned by  $x_1, \dots, x_m$ .*

PROOF. With  $x_1, \dots, x_m$  as stated, each  $-x_j$  is in  $\text{dom } \sigma$ . Let  $x := \sum_{j=1}^m t_j x_j$  be an arbitrary linear combination of  $x_1, \dots, x_m$ ; we must prove that  $\sigma(x) = \sum_{j=1}^m t_j \sigma(x_j)$ . Set

$$J_1 := \{j : t_j > 0\}, \quad J_2 := \{j : t_j < 0\}$$

and obtain (as usual,  $\sum \emptyset = 0$ ):

$$\begin{aligned} \sigma(x) &= \sigma\left(\sum_{J_1} t_j x_j + \sum_{J_2} (-t_j)(-x_j)\right) \\ &\leq \sum_{J_1} t_j \sigma(x_j) + \sum_{J_2} (-t_j) \sigma(-x_j) && [\text{from (1.1.4)}] \\ &= \sum_{J_1} t_j \sigma(x_j) + \sum_{J_2} t_j \sigma(x_j) = \sum_{j=1}^m t_j \sigma(x_j) && [\text{from (1.1.7)}] \\ &= -\sum_{J_1} t_j \sigma(-x_j) - \sum_{J_2} (-t_j) \sigma(x_j) && [\text{from (1.1.7)}] \\ &\leq -\sigma\left(-\sum_{j=1}^m t_j x_j\right) && [\text{from (1.1.4)}] \\ &= -\sigma(-x) \leq \sigma(x). && [\text{from (1.1.6)}] \end{aligned}$$

In summary, we have proved

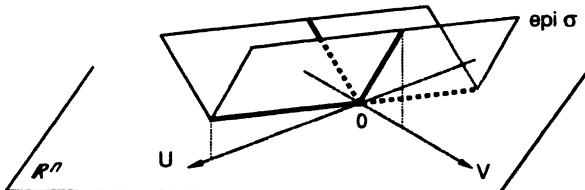
$$\sigma(x) \leq \sum_{j=1}^m t_j \sigma(x_j) \leq -\sigma(-x) \leq \sigma(x).$$

□

Thanks to this result, we are entitled to define

$$U := \{x \in \mathbb{R}^n : \sigma(x) + \sigma(-x) = 0\} \quad (1.1.8)$$

which is a subspace of  $\mathbb{R}^n$ : the subspace in which  $\sigma$  is *linear*. Note that  $U$  nonempty corresponds to  $\sigma(0) = 0$  (even if  $U$  reduces to  $\{0\}$ ).



**Fig. 1.1.2.** Subspace of linearity of a sublinear function

What is interesting in this concept is its geometric interpretation. If  $V$  is another subspace such that  $U \cap V = \{0\}$ , there holds by definition

$$\sigma(x) + \sigma(-x) > 0 \quad \text{for all } 0 \neq x \in V.$$

This means that, if  $0 \neq d \in V$ ,  $\sigma$  is “V-shaped” along  $d$ : for  $t > 0$ ,

$$\sigma(td) = \alpha t \quad \text{and} \quad \sigma(-td) = \beta t$$

for some  $\alpha$  and  $\beta$  in  $\mathbb{R} \cup \{+\infty\}$  such that  $\alpha + \beta > 0$ ; whereas  $\alpha$  and  $\beta$  would be 0 if  $d$  were in  $U$ . See Fig. 1.1.2 for an illustration. For  $d$  of norm 1, the number  $\alpha + \beta$  above could be

called the “lack of linearity” of  $\sigma$  along  $d$ : when restricted to the line  $d$ , the graph of  $\sigma$  makes an angle; when finite, the number  $\alpha + \beta$  measures how acute this angle is.

Figure 1.1.2 suggests that  $\text{gr } \sigma$  is a hyperplane not only in  $U$ , but also in the translations of  $U$ : the restriction of  $\sigma$  to  $\{y\} + U$  is affine, for any fixed  $y$ . This comes from the next result.

**Proposition 1.1.7** *Let  $\sigma$  be sublinear. If  $x \in U$ , i.e. if*

$$\sigma(x) + \sigma(-x) = 0, \quad (1.1.9)$$

*then there holds*

$$\sigma(x+y) = \sigma(x) + \sigma(y) \quad \text{for all } y \in \mathbb{R}^n. \quad (1.1.10)$$

PROOF. In view of subadditivity, we just have to prove “ $\geq$ ” in (1.1.10). Start from the identity  $y = x + y - x$ ; apply successively subadditivity and (1.1.9) to obtain

$$\sigma(y) \leq \sigma(x+y) + \sigma(-x) = \sigma(x+y) - \sigma(x).$$

□

## 1.2 Some Examples

We start with some simple situations. If  $K$  is a nonempty convex cone, its indicator function

$$I_K(x) := \begin{cases} 0 & \text{if } x \in K, \\ +\infty & \text{if not} \end{cases}$$

is clearly sublinear. In  $\mathbb{R}^n \times \mathbb{R}$ , the epigraph  $\text{epi } I_K$  is made up of all the copies of  $K$ , shifted upwards. Likewise, a distance function

$$d_K(x) := \inf \{\|y-x\| : y \in K\}$$

is also sublinear: nothing in the picture is essentially changed when both  $x$  and  $y$  are multiplied by  $t > 0$ . Another example is the function from  $\mathbb{R}^2$  to  $\mathbb{R} \cup \{+\infty\}$

$$\sigma(x) = \sigma(\xi, \eta) := \begin{cases} -2\sqrt{\xi\eta} & \text{if } \xi, \eta \geq 0 \\ +\infty & \text{if not.} \end{cases}$$

Its positive homogeneity is clear, its convexity is not particularly difficult to check (see Example IV.4.3.3), it is therefore sublinear. A good exercise is to try to visualize its epigraph.

**Example 1.2.1** Let  $f \in \text{Conv } \mathbb{R}^n$ ; its perspective  $\tilde{f}$  of §IV.2.2, which is convex, is clearly positively homogeneous (from  $\mathbb{R}^{n+1}$  to  $\mathbb{R} \cup \{+\infty\}$ ); it is an important instance of sublinear function. For example, in  $\mathbb{R}^2$

$$\tilde{f}(u, \xi) := \begin{cases} \frac{1}{2}\xi^2/u & \text{if } u > 0, \\ +\infty & \text{if not} \end{cases} \quad (1.2.1)$$

is the perspective of  $\xi \mapsto f(\xi) = \frac{1}{2}\xi^2$ .

Note that  $\tilde{f}(0, 0) = +\infty$ . The closure of  $\tilde{f}$  can be computed with the help of Example IV.3.2.4: clearly enough, the asymptotic function of  $f$  is  $I_{\{0\}}$ . Hence  $(\text{cl } \tilde{f})(0, 0) = 0$ , while  $\tilde{f}$  coincides with its closure everywhere else. □

**Example 1.2.2 (Norms)** We recall that a norm  $\|\cdot\|$  on  $\mathbb{R}^n$  is a function from  $\mathbb{R}^n$  to  $[0, +\infty[$  satisfying the following properties:

- (i)  $\|x\| = 0$  if and only if  $x = 0$ ;
- (ii)  $\|tx\| = |t| \|x\|$  for all  $x \in \mathbb{R}^n$  and  $t \in \mathbb{R}$ ;
- (iii)  $\|x_1 + x_2\| \leq \|x_1\| + \|x_2\|$  for all  $(x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^n$ .

Clearly,  $\|\cdot\|$  is a positive (except at 0) and finite sublinear function which, moreover, is symmetric i.e.  $\|-x\| = \|x\|$  for all  $x$ . It is linear on no line: the subspace  $U$  of (1.1.8) is reduced to  $\{0\}$ .

Conversely, if  $\sigma$  is a sublinear function from  $\mathbb{R}^n$  into  $[0, +\infty[$  which is linear on no line, i.e. such that

$$\sigma(x) + \sigma(-x) > 0 \quad \text{for all } x \neq 0,$$

then  $\|x\| := \max\{\sigma(x), \sigma(-x)\}$  is a norm on  $\mathbb{R}^n$ .  $\square$

**Example 1.2.3 (Quadratic Semi-Norms)** Take a symmetric positive semi-definite operator  $Q$  from  $\mathbb{R}^n$  to  $\mathbb{R}^n$  and define

$$f(x) := \sqrt{\langle Qx, x \rangle} \quad \text{for all } x \in \mathbb{R}^n.$$

Convexity of  $f$  (i.e. its subadditivity, i.e. the Cauchy-Schwarz inequality) is not so easy to prove directly. Consider, however, the convex set

$$E_Q := \{x \in \mathbb{R}^n : \langle Qx, x \rangle \leq 1\}.$$

Then  $f$  can be obtained as follows:

$$\begin{aligned} f(x) &= \inf \{\lambda > 0 : \langle Qx, x \rangle \leq \lambda^2\} \\ &= \inf \{\lambda > 0 : \langle Q\frac{x}{\lambda}, \frac{x}{\lambda} \rangle \leq 1\} \\ &= \inf \{\lambda > 0 : \frac{x}{\lambda} \in E_Q\} \end{aligned}$$

and we will see below that this establishes convexity – hence sublinearity – of  $f$ .

Observe in passing that  $E_Q$  is the sublevel-set at level 1 of both  $f$  and  $f^2 = \langle Q \cdot, \cdot \rangle$ . Decompose the space as  $\mathbb{R}^n = \text{Ker } Q \oplus \text{Im } Q$ : the intersection of  $E_Q$  with  $\text{Im } Q$  is an ellipsoid centered at the origin, say  $\tilde{E}_Q$ . The entire  $E_Q$  is the cylinder  $\tilde{E}_Q + \text{Ker } Q$ , whose asymptotic cone is just the subspace  $\text{Ker } Q$ . If and only if  $\text{Ker } Q = \{0\}$ , i.e.  $Q$  is positive definite, is  $E_Q$  compact, namely an ellipsoid. On the other hand,  $f$  is finite, nonnegative, symmetric because  $E_Q$  has center 0; and  $f$  is zero on the asymptotic cone  $\text{Ker } Q$  of  $E_Q$ . Theorem 1.2.5 below establishes the convexity of  $f$ , which is therefore a semi-norm, actually a norm if  $Q$  is positive definite.  $\square$

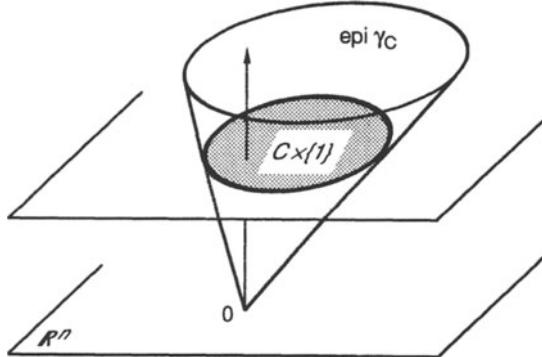
The mapping  $E_Q \mapsto f$ , introduced in Example 1.2.3, is important in the context of sublinear functions; let us put it in perspective.

**Definition 1.2.4** Let  $C$  be a closed convex set containing the origin. The function  $\gamma_C$  defined by

$$\gamma_C(x) := \inf \{\lambda > 0 : x \in \lambda C\} \tag{1.2.2}$$

is called the *gauge* of  $C$ . As usual, we set  $\gamma_C(x) := +\infty$  if  $x \notin \lambda C$  for no  $\lambda > 0$ .  $\square$

Geometrically,  $\gamma_C$  can be obtained as follows: shift  $C$  ( $\subset \mathbb{R}^n$ ) in the hyperplane  $\mathbb{R}^n \times \{1\}$  of the graph-space  $\mathbb{R}^n \times \mathbb{R}$  (by contrast to a perspective-function, the present shift is vertical, along the axis of function-values). Then the epigraph of  $\gamma_C$  is the cone generated by this shifted copy of  $C$ ; see Fig. 1.2.1.



**Fig. 1.2.1.** The epigraph of a gauge

The next result summarizes the main properties of a gauge. Each statement should be read with Fig. 1.2.1 in mind, even though the picture is slightly misleading, due to closure problems.

**Theorem 1.2.5** *Let  $C$  be a closed convex set containing the origin. Then*

- (i) *its gauge  $\gamma_C$  is a nonnegative closed sublinear function;*
- (ii)  *$\gamma_C$  is finite everywhere if and only if 0 lies in the interior of  $C$ ;*
- (iii)  *$C_\infty$  being the asymptotic cone of  $C$ ,*

$$\begin{aligned}\{x \in \mathbb{R}^n : \gamma_C(x) \leq r\} &= rC \quad \text{for all } r > 0, \\ \{x \in \mathbb{R}^n : \gamma_C(x) = 0\} &= C_\infty.\end{aligned}$$

**PROOF.** [(i) and (iii)] Nonnegativity and positive homogeneity are obvious from the definition of  $\gamma_C$ ; also,  $\gamma_C(0) = 0$  because  $0 \in C$ . We prove convexity via a geometric interpretation of (1.2.2). Let

$$K_C := \text{cone}(C \times \{1\}) = \{(\lambda c, \lambda) \in \mathbb{R}^n \times \mathbb{R} : c \in C, \lambda \geq 0\}$$

be the convex conical hull of  $C \times \{1\} \subset \mathbb{R}^n \times \mathbb{R}$ . It is convex (beware that  $K_C$  need not be closed) and  $\gamma_C$  is clearly given by

$$\gamma_C(x) = \inf \{\lambda : (x, \lambda) \in K_C\}.$$

Thus,  $\gamma_C$  is the lower-bound function of §IV.1.3(g), constructed on the convex set  $K_C$ ; this establishes the convexity of  $\gamma_C$ , hence its sublinearity.

Now we prove

$$\{x \in \mathbb{R}^n : \gamma_C(x) \leq 1\} = C. \tag{1.2.3}$$

This will imply the first part in (iii), thanks to positive homogeneity. Then the second part will follow because of (III.2.2.2):

$$C_\infty = \cap\{rC : r > 0\}$$

and closedness of  $\gamma_C$  will also result from (iii) via Proposition IV.1.2.2.

So, to prove (1.2.3), observe first that  $x \in C$  implies from (1.2.2) that certainly  $\gamma_C(x) \leq 1$ . Conversely, let  $x$  be such that  $\gamma_C(x) \leq 1$ ; we must prove that  $x \in C$ . For this we prove that  $x_k := (1 - 1/k)x \in C$  for  $k = 1, 2, \dots$  (and then, the desired property will come from the closedness of  $C$ ). By positive homogeneity,  $\gamma_C(x_k) = (1 - 1/k)\gamma_C(x) < 1$ , so there is  $\lambda_k \in ]0, 1[$  such that  $x_k \in \lambda_k C$ , or equivalently  $x_k/\lambda_k \in C$ . Because  $C$  is convex and contains the origin,  $\lambda_k(x_k/\lambda_k) + (1 - \lambda_k)0 = x_k$  is in  $C$ , which is what we want.

[(ii)] Assume  $0 \in \text{int } C$ . There is  $\varepsilon > 0$  such that for all  $x \neq 0$ ,  $x_\varepsilon := \varepsilon x / \|x\| \in C$ ; hence  $\gamma_C(x_\varepsilon) \leq 1$  because of (1.2.3). We deduce by positive homogeneity

$$\gamma_C(x) = \frac{\|x\|}{\varepsilon} \gamma_C(x_\varepsilon) \leq \frac{\|x\|}{\varepsilon};$$

this inequality actually holds for all  $x \in \mathbb{R}^n$  ( $\gamma_C(0) = 0$ ) and  $\gamma_C$  is a finite function.

Conversely, suppose  $\gamma_C$  is finite everywhere. By continuity (Theorem IV.3.1.2),  $\gamma_C$  has an upper bound  $L > 0$  on the unit ball:

$$\|x\| \leq 1 \implies \gamma_C(x) \leq L \implies x \in LC,$$

where the last implication comes from (iii). In other words,  $B(0, 1/L) \subset C$ . □

Since  $\gamma_C$  is the lower-bound function of the cone  $K_C$  ( $= K_C + \{0\} \times \mathbb{R}^+$ ) of Fig. 1.2.1, we know from (IV.1.3.6) that

$$K_C \subset \text{epi } \gamma_C \subset \text{cl } K_C;$$

but  $\gamma_C$  has a closed epigraph, therefore

$$\text{epi } \gamma_C = \text{cl } K_C = \overline{\text{cone}}(C \times \{1\}). \quad (1.2.4)$$

Since  $C_\infty = \{0\}$  if and only if  $C$  is compact (Proposition III.2.2.3), we obtain another consequence of (iii):

**Corollary 1.2.6** *C is compact if and only if  $\gamma_C(x) > 0$  for all  $x \neq 0$ .* □

**Example 1.2.7** The quadratic semi-norms of Example 1.2.3 can be generalized: let  $f \in \text{Conv } \mathbb{R}^n$  have nonnegative values and be positively homogeneous of degree 2, i.e.

$$0 \leq f(tx) = t^2 f(x) \quad \text{for all } x \in \mathbb{R}^n \text{ and all } t > 0.$$

Then,  $\sqrt{f}$  is convex; in fact

$$\begin{aligned}\sqrt{f}(x) &= \inf \{\lambda > 0 : \sqrt{f(x)} \leq \lambda\} \\ &= \inf \{\lambda > 0 : f(x) \leq \lambda^2\} \\ &= \inf \{\lambda > 0 : \frac{x}{\lambda} \in S_1(f)\},\end{aligned}$$

which displays the sublevel-set

$$S_1(f) = \{x \in \mathbb{R}^n : f(x) \leq 1\} =: C.$$

In other words,  $\sqrt{f}$  is the gauge of a closed convex set  $C$  containing the origin.  $\square$

Gauges are examples of sublinear functions which are closed. This is not the case of all sublinear functions: see the function  $\tilde{f}$  of (1.2.1); another example in  $\mathbb{R}^2$  is

$$h(\xi, \eta) := \begin{cases} 0 & \text{if } \eta > 0, \\ |\xi| & \text{if } \eta = 0, \\ +\infty & \text{if } \eta < 0. \end{cases}$$

By taking the closure, or lower semi-continuous hull, of a sublinear function  $\sigma$ , we get a new function defined by

$$\text{cl } \sigma(x) := \liminf_{x' \rightarrow x} \sigma(x') \quad (1.2.5)$$

which is (i) closed by construction, (ii) convex (Proposition IV.1.2.6) and (iii) positively homogeneous, as is immediately seen from (1.2.5). For example, to close the above  $h$ , one must set  $h(\xi, 0) = 0$  for all  $\xi$ . We retain from this observation that, when we close a sublinear function, we obtain a new function which is closed, of course, but which inherits sublinearity. The subclass of sublinear functions that are also closed is extremely important, particularly for minimization; in fact most of our study will be restricted to these.

Note that, for a closed sublinear function  $\sigma$ ,

$$\sigma(0) \leq \lim_{t \downarrow 0} \sigma(tx) = 0 \quad \text{for all } x \in \text{dom } \sigma,$$

so certainly  $\sigma(0) = 0$ ; otherwise,  $\text{dom } \sigma$  would be empty, a situation that we reject from our definitions. Another observation is that a closed sublinear function  $\sigma$  coincides with its asymptotic function:

$$\sigma'_\infty = \sigma \quad \text{if } \sigma \text{ is closed and sublinear}$$

(take  $x_0 = 0$  in the definition of Proposition IV.3.2.2). In particular, if  $\sigma$  is finite everywhere, then Proposition IV.3.2.7 tells us that it is Lipschitzian, and its best Lipschitz constant is

$$\sup \{\sigma(d) : \|d\| = 1\}. \quad (1.2.6)$$

### 1.3 The Convex Cone of All Closed Sublinear Functions

Similarly to convex functions, sublinear functions, closed or not, can be combined to give new sublinear functions.

#### Proposition 1.3.1

- (i) If  $\sigma_1$  and  $\sigma_2$  are [closed] sublinear and  $t_1, t_2$  are positive numbers, then  $\sigma := t_1\sigma_1 + t_2\sigma_2$  is [closed] sublinear, if not identically  $+\infty$ .
- (ii) If  $\{\sigma_j\}_{j \in J}$  is a family of [closed] sublinear functions, then  $\sigma := \sup_{j \in J} \sigma_j$  is [closed] sublinear, if not identically  $+\infty$ .

PROOF. Concerning convexity and closedness, everything is known from §IV.2. Note in passing that a closed sublinear function is zero (hence finite) at zero. As for positive homogeneity, it is straightforward.  $\square$

**Proposition 1.3.2** Let  $\{\sigma_j\}_{j \in J}$  be a family of sublinear functions all minorized by some linear function. Then

- (i)  $\sigma := \text{co}(\inf_{j \in J} \sigma_j)$  is sublinear.
- (ii) If  $J = \{1, \dots, m\}$  is a finite set, we obtain the infimal convolution

$$\text{co min}\{\sigma_1, \dots, \sigma_m\} = \sigma_1 \downarrow \dots \downarrow \sigma_m.$$

PROOF. [(i)] Once again, the only thing to prove for (i) is positive homogeneity. Actually, it suffices to multiply  $x$  and each  $x_j$  by  $t > 0$  in a formula giving  $\text{co}(\inf_j \sigma_j)(x)$ , say (IV.2.5.4).

[(ii)] By definition, computing  $\text{co}(\min_j \sigma_j)(x)$  amounts to solving the minimization problem in the  $m$  couples of variables  $(x_j, \alpha_j) \in \text{dom } \sigma_j \times \mathbb{R}$

$$\left| \begin{array}{l} \inf \sum_{j=1}^m \alpha_j \sigma_j(x_j) \\ \sum_{j=1}^m \alpha_j = 1, \quad \sum_{j=1}^m \alpha_j x_j = x \end{array} \right. \quad (1.3.1)$$

In view of positive homogeneity, the variables  $\alpha_j$  play no role by themselves: the relevant variables are actually the products  $\alpha_j x_j$  and (1.3.1) can be written – denoting  $\alpha_j x_j$  again by  $x_j$ :

$$\inf \left\{ \sum_{j=1}^m \sigma_j(x_j) : \sum_{j=1}^m x_j = x \right\}.$$

We recognize the infimal convolution of the  $\sigma_j$ 's.  $\square$

From Proposition 1.3.1(i), the collection of all closed sublinear functions has an *algebraic* structure: it is a convex cone contained in  $\text{Conv } \mathbb{R}^n$ . It contains another convex cone, namely the collection of finite sublinear functions.

A *topological* structure can be defined on the latter cone. In linear analysis, one defines the Euclidean distance between two linear forms  $\ell_1 = \langle s_1, \cdot \rangle$  and  $\ell_2 = \langle s_2, \cdot \rangle$ :

$$\|\ell_1 - \ell_2\| := \|s_1 - s_2\| = \max_{\|x\| \leq 1} |\ell_1(x) - \ell_2(x)|.$$

A distance can also be defined on the convex cone of *everywhere finite* sublinear functions (the extended-valued case is somewhat more delicate), which of course contains the vector space of linear forms.

**Theorem 1.3.3** For  $\sigma_1$  and  $\sigma_2$  in the set  $\Phi$  of sublinear functions that are finite everywhere, define

$$\Delta(\sigma_1, \sigma_2) := \max_{\|x\| \leq 1} |\sigma_1(x) - \sigma_2(x)|. \quad (1.3.2)$$

Then  $\Delta$  is a distance on  $\Phi$ .

PROOF. Clearly  $\Delta(\sigma_1, \sigma_2) < +\infty$  and  $\Delta(\sigma_1, \sigma_2) = \Delta(\sigma_2, \sigma_1)$ . Now positive homogeneity of  $\sigma_1$  and  $\sigma_2$  gives for all  $x \neq 0$

$$\begin{aligned} |\sigma_1(x) - \sigma_2(x)| &= \|x\| \left| \sigma_1\left(\frac{x}{\|x\|}\right) - \sigma_2\left(\frac{x}{\|x\|}\right) \right| \\ &\leq \|x\| \max_{\|u\|=1} |\sigma_1(u) - \sigma_2(u)| \\ &\leq \|x\| \Delta(\sigma_1, \sigma_2). \end{aligned}$$

In addition,  $\sigma_1(0) = \sigma_2(0) = 0$ , so

$$|\sigma_1(x) - \sigma_2(x)| \leq \|x\| \Delta(\sigma_1, \sigma_2) \quad \text{for all } x \in \mathbb{R}^n$$

and  $\Delta(\sigma_1, \sigma_2) = 0$  if and only if  $\sigma_1 = \sigma_2$ .

As for the triangle inequality, we have for arbitrary  $\sigma_1, \sigma_2, \sigma_3$  in  $\Phi$

$$|\sigma_1(x) - \sigma_3(x)| \leq |\sigma_1(x) - \sigma_2(x)| + |\sigma_2(x) - \sigma_3(x)| \quad \text{for all } x \in \mathbb{R}^n,$$

so there holds

$$\begin{aligned} \Delta(\sigma_1, \sigma_3) &\leq \max_{\|x\| \leq 1} [|\sigma_1(x) - \sigma_2(x)| + |\sigma_2(x) - \sigma_3(x)|] \\ &\leq \max_{\|x\| \leq 1} |\sigma_1(x) - \sigma_2(x)| + \max_{\|x\| \leq 1} |\sigma_2(x) - \sigma_3(x)| \end{aligned}$$

which is the required inequality.  $\square$

The index-set in (1.3.2) can be replaced by the unit sphere  $\|x\| = 1$ , just as in (1.2.6); and the distance between an arbitrary  $\sigma \in \Phi$  and the zero-function (which is in  $\Phi$ ) is just the value (1.2.6). The function  $\Delta(\cdot, 0)$  acts like a *norm* on the convex cone  $\Phi$ .

**Example 1.3.4** Consider  $\|\cdot\|_1$  and  $\|\cdot\|_\infty$ , the  $\ell_1$ - and  $\ell_\infty$ -norms on  $\mathbb{R}^n$ . They are finite sublinear (Example 1.2.2) and there holds

$$\Delta(\|\cdot\|_1, \|\cdot\|_\infty) = \frac{n-1}{\sqrt{n}}.$$

To accept this formula, consider that, for symmetry reasons, the maximum in the definition (1.3.2) of  $\Delta$  is achieved at  $x = (1/\sqrt{n}, \dots, 1/\sqrt{n})$ .  $\square$

The convergence associated with this new distance function turns out to be the natural one:

**Theorem 1.3.5** Let  $\{\sigma_k\}$  be a sequence of finite sublinear functions and let  $\sigma$  be a finite function. Then the following are equivalent when  $k \rightarrow +\infty$ :

- (i)  $\{\sigma_k\}$  converges pointwise to  $\sigma$ ;

- (ii)  $\{\sigma_k\}$  converges to  $\sigma$  uniformly on each compact set of  $\mathbb{R}^n$ ;
- (iii)  $\Delta(\sigma_k, \sigma) \rightarrow 0$ .

PROOF. First, the (finite) function  $\sigma$  is of course sublinear whenever it is the pointwise limit of sublinear functions. The equivalence between (i) and (ii) comes from the general Theorem IV.3.1.5 on the convergence of convex functions.

Now, (ii) clearly implies (iii). Conversely  $\Delta(\sigma_k, \sigma) \rightarrow 0$  is the uniform convergence on the unit ball, hence on any ball of radius  $L > 0$  (the maximand in (1.3.2) is positively homogeneous), hence on any compact set.  $\square$

## 2 The Support Function of a Nonempty Set

### 2.1 Definitions, Interpretations

**Definition 2.1.1** Let  $S$  be a nonempty set in  $\mathbb{R}^n$ . The function  $\sigma_S : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$  defined by

$$\mathbb{R}^n \ni x \mapsto \sigma_S(x) := \sup \{\langle s, x \rangle : s \in S\} \quad (2.1.1)$$

is called the *support function* of  $S$ .  $\square$

For a given  $S$ , the support function is therefore attached to the scalar product  $\langle \cdot, \cdot \rangle$ : in (2.1.1), the space where  $s$  runs and the space where  $\sigma_S$  acts are dual to each other. It follows, for example, that if the scalar product is changed,  $S$  remaining the same,  $\sigma_S$  is changed.

The supremum in (2.1.1) may be finite or infinite, achieved on  $S$  or not. In this context,  $S$  can be interpreted as an index set:  $\sigma_S(\cdot)$  is the supremum of the collection of linear forms  $\langle s, \cdot \rangle$  over  $S$ . We obtain immediately:

**Proposition 2.1.2** *A support function is closed and sublinear.*

PROOF. This results from Proposition 1.3.1(ii) (a linear form is closed and convex!). Observe in particular that a support function is null (hence  $< +\infty$ ) at the origin.  $\square$

The domain of  $\sigma_S$  is a convex cone, closed or not. Actually,  $x \in \text{dom } \sigma_S$  means that, for some  $r := \sigma_S(x)$ :

$$S \subset \{s \in \mathbb{R}^n : \langle s, x \rangle \leq r\} \quad (2.1.2)$$

i.e.  $S$  is contained in a closed half-space “opposite” to  $x$ .

**Proposition 2.1.3** *The support function of  $S$  is finite everywhere if and only if  $S$  is bounded.*

PROOF. Let  $S$  be bounded, say  $S \subset B(0, L)$  for some  $L > 0$ . Then

$$\langle s, x \rangle \leq \|s\| \|x\| \leq L \|x\| \quad \text{for all } s \in S,$$

which implies  $\sigma_S(x) \leq L \|x\|$  for all  $x \in \mathbb{R}^n$ .

Conversely, finiteness of the convex  $\sigma_S$  implies its continuity on the whole space (Theorem IV.3.1.2), hence its local boundedness: for some  $L$ ,

$$\langle s, x \rangle \leq \sigma_S(x) \leq L \quad \text{for all } (s, x) \in S \times B(0, 1).$$

If  $s \neq 0$ , we can take  $x = s/\|s\|$  in the above relation, which implies  $\|s\| \leq L$ .  $\square$

Observing that

$$-\sigma_S(-x) = -\sup_{s \in S} [-\langle s, x \rangle] = \inf_{s \in S} \langle s, x \rangle,$$

the number  $\sigma_S(x) + \sigma_S(-x)$  of (1.1.6) is particularly interesting here:

**Definition 2.1.4** The *breadth* of the nonempty set  $S$  along  $x \neq 0$  is

$$\sigma_S(x) + \sigma_S(-x) = \sup_{s \in S} \langle s, x \rangle - \inf_{s \in S} \langle s, x \rangle,$$

a number in  $[0, +\infty]$ . It is 0 if and only if  $S$  lies entirely in some affine hyperplane orthogonal to  $x$ ; such a hyperplane is expressed as

$$\{y \in \mathbb{R}^n : \langle y, x \rangle = \sigma_S(x)\},$$

which in particular contains  $S$ . The intersection of all these hyperplanes is just the affine hull of  $S$ .  $\square$

If  $x$  has norm 1 and is interpreted as a *direction*, the breadth of  $S$  measures how “thick”  $S$  is along  $x$ : it is the distance between the two hyperplanes orthogonal to  $x$  and “squeezing”  $S$ . This observation calls for a more general comment: a sublinear function  $x \mapsto \sigma(x)$  being positively homogeneous, the norm of its argument  $x$  has little importance. This argument should always be thought of as an *oriented direction*, i.e. a normalized vector of  $\mathbb{R}^n$ . Accordingly, we will generally use from now on the notation  $\sigma(d)$ , more suggestive for a support function than  $\sigma(x)$ .

Here, we give two geometric constructions which help interpreting a support function.

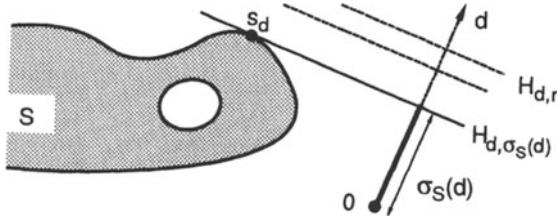
**Interpretation 2.1.5 (Construction in  $\mathbb{R}^n$ )** Given  $S \subset \mathbb{R}^n$  and  $d \neq 0$ , consider for each  $r \in \mathbb{R}$  the closed half-space alluded to in (2.1.2):

$$H_{d,r}^- := \{z \in \mathbb{R}^n : \langle z, d \rangle \leq r\}. \tag{2.1.3}$$

If (2.1.2) holds, we can find  $r$  large enough so that  $S \subset H_{d,r}^-$ . The value  $\sigma_S(d)$  is the smallest of those  $r$ : decreasing  $r$  as much as possible while keeping  $S$  in  $H_{d,r}^-$  means “leaning” onto  $S$  the affine hyperplane

$$H_{d,r} := \{z \in \mathbb{R}^n : \langle z, d \rangle = r\}.$$

See Fig. 2.1.1 for an illustration.



**Fig. 2.1.1.** Supporting hyperplanes and support functions

If (2.1.2) does not hold, however, this operation is impossible:  $S$  is “unbounded in the oriented direction”  $d$  and  $\sigma_S(d) = +\infty$ . Take for example in  $\mathbb{R}^2$

$$S := \{(\xi, 0) : \xi \geq 0\}.$$

For  $d = (1, 1)$  say (and assuming that  $\langle \cdot, \cdot \rangle$  is the usual dot-product), no closed half-space of the form (2.1.3) can contain  $S$ , even if  $r$  is increased to  $+\infty$ .

If  $S$  is compact, the supremum of the linear form  $\langle \cdot, d \rangle$  is achieved on  $S$ , no matter how  $d$  is chosen. This means that, somewhere on the hyperplane  $H_{d,\sigma_S(d)}$  there is some  $s_d$  which is also in  $S$ , actually a boundary point of  $S$ .  $\square$

Figure 2.1.1 suggests (and Proposition 2.2.1 below confirms) that the support functions of  $S$  and of  $\overline{\text{co}} S$  coincide. Note also that the distance from the origin 0 to the “optimal” hyperplane  $H_{d,\sigma_S(d)}$  is  $|\sigma_S(d/\|d\|)|$ . This is easily confirmed: project the origin onto  $H_{d,\sigma_S(d)}$  to obtain the vector  $t^*d$  such that  $\langle d, t^*d \rangle = \sigma_S(d)$ . Then the distance from 0 to  $H_{d,\sigma_S(d)}$  is  $\|t^*d\|$ .

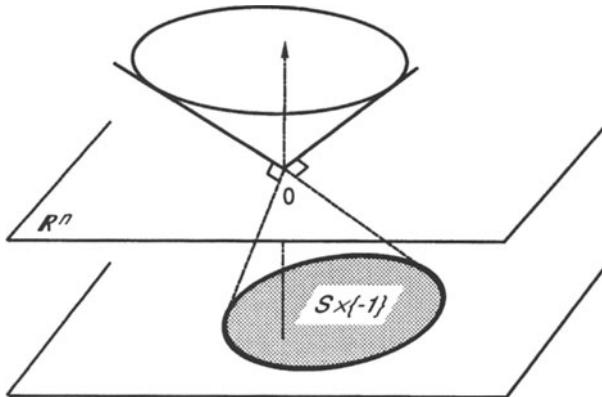
**Interpretation 2.1.6 (Construction in  $\mathbb{R}^{n+1}$ )** In the graph-space  $\mathbb{R}^n \times \mathbb{R}$ , we shift  $S$  down to  $\mathbb{R}^n \times \{-1\}$  and consider the convex conical hull  $K_S$  of this shifted copy of  $S$ . Then the polar cone  $(K_S)^\circ$  of  $K_S$  is nothing else than the epigraph of  $\sigma_S$ . Indeed

$$K_S = \mathbb{R}^+ \text{co}(S \times \{-1\}) = \text{co} [\mathbb{R}^+(S \times \{-1\})],$$

so that

$$\begin{aligned} (K_S)^\circ &= \{(d, r) : t\langle s, d \rangle - tr \leq 0 \text{ for all } s \in S \text{ and } t > 0\} \\ &= \{(d, r) : \langle s, d \rangle \leq r \text{ for all } s \in S\} \\ &= \{(d, r) : \sup_{s \in S} \langle s, d \rangle \leq r\} = \text{epi } \sigma_S. \end{aligned}$$

This is illustrated on Fig. 2.1.2. We have intentionally chosen a case with  $0 \in S$ . It implies  $\sigma_S(d) \geq 0$  for all  $d$ , as is obvious just from its definition (2.1.1). This property is fundamental in optimization (and frankly, the picture is then a lot easier to draw!).  $\square$



**Fig. 2.1.2.** The epigraph of a support function

## 2.2 Basic Properties

First, we list some properties of support functions that are directly derived from their definition.

**Proposition 2.2.1** *For  $S \subset \mathbb{R}^n$  nonempty, there holds  $\sigma_S = \sigma_{\text{cl } S} = \sigma_{\text{co } S}$ ; whence*

$$\sigma_S = \sigma_{\text{co } S}. \quad (2.2.1)$$

PROOF. The continuity [resp. linearity, hence convexity] of the function  $\langle s, \cdot \rangle$ , which is maximized over  $S$ , implies that  $\sigma_S = \sigma_{\text{cl } S}$  [resp.  $\sigma_S = \sigma_{\text{co } S}$ ]. Knowing that  $\text{co } S = \text{cl co } S$  (Proposition III.1.4.2), (2.2.1) follows immediately.  $\square$

This result is of utmost importance: it says that the concept of support function *does not distinguish* a set  $S$  from its closed convex hull. Thus, when dealing with support functions, it makes no difference if we restrict ourselves to the case of closed convex sets.

As a result of (2.1.1) and (2.2.1), we can write

$$s \in \text{co } S \implies [\langle s, d \rangle \leq \sigma_S(d) \text{ for all } d \in \mathbb{R}^n].$$

Now, what about the converse? Can it be that the above (infinite) set of inequalities still holds if  $s$  is not in  $\text{co } S$ ? The answer is no:

**Theorem 2.2.2** *For the nonempty  $S \subset \mathbb{R}^n$  and its support function  $\sigma_S$ , there holds*

$$s \in \text{co } S \iff [\langle s, d \rangle \leq \sigma_S(d) \text{ for all } d \in X], \quad (2.2.2)$$

where the set  $X$  can be indifferently taken as: the whole of  $\mathbb{R}^n$ , the unit ball  $B(0, 1)$  or its boundary  $\tilde{B}$ , or  $\text{dom } \sigma_S$ .

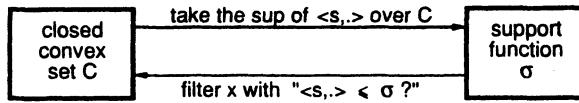
PROOF. First, the equivalence between all the choices for  $X$  is clear enough; in particular due to positive homogeneity. Because “ $\Rightarrow$ ” is Proposition 2.2.1, we have to prove “ $\Leftarrow$ ” only, with  $X = \mathbb{R}^n$  say.

Suppose that  $s \notin \overline{\text{co}} S$ . Then  $\{s\}$  and  $\overline{\text{co}} S$  can be strictly separated (Theorem III.4.1.1): there exists  $d_0 \in \mathbb{R}^n$  such that

$$\langle s, d_0 \rangle > \sup \{ \langle s', d_0 \rangle : s' \in \overline{\text{co}} S \} = \sigma_S(d_0),$$

where the last equality is (2.2.1). Our result is proved by contradiction.  $\square$

As a result, a closed convex set is completely determined by its support function: between the classes of closed convex sets and of support functions, there is a correspondence which is bijective, as illustrated on Fig. 2.2.1.



**Fig. 2.2.1.** Correspondence between closed convex sets and support functions

Thus, whether a given point  $s$  belongs to a given closed convex set  $S$  can be checked with the help of (2.2.2), which holds as an equivalence. Actually, more can be said: the support function *filters* the interior, the relative interior and the affine hull of a closed convex set.

This property is best understood with Fig. 1.1.2 in mind. Let  $V$  be the subspace parallel to  $\text{aff } S$ , and  $U := V^\perp$ . Indeed,  $U$  is just given in (1.1.8) with  $\sigma = \sigma_S$ :  $U$  [resp.  $V$ ] can be viewed either as the subspace where the sublinear  $\sigma_S$  is linear [resp. kinky], or where the supported set  $S$  is flat [resp. thick]. When drawn in the geometric space of convex sets, Fig. 1.1.2 becomes Fig. 2.2.2, which is very helpful to follow the next proof.

**Theorem 2.2.3** *Let  $S$  be a nonempty closed convex set in  $\mathbb{R}^n$ . Then*

(i)  $s \in \text{aff } S$  if and only if

$$\langle s, d \rangle = \sigma_S(d) \quad \text{for all } d \text{ with } \sigma_S(d) + \sigma_S(-d) = 0; \quad (2.2.3)$$

(ii)  $s \in \text{ri } S$  if and only if

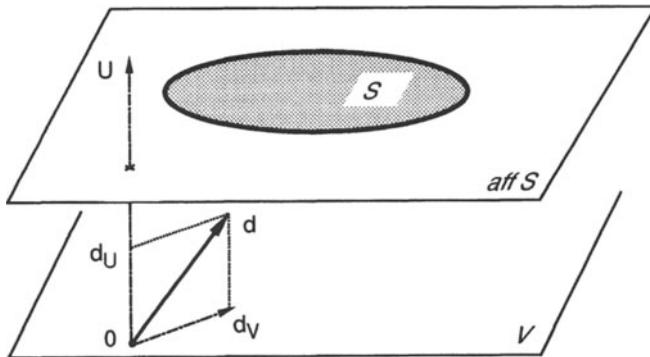
$$\langle s, d \rangle < \sigma_S(d) \quad \text{for all } d \text{ with } \sigma_S(d) + \sigma_S(-d) > 0; \quad (2.2.4)$$

(iii) in particular,  $s \in \text{int } S$  if and only if

$$\langle s, d \rangle < \sigma_S(d) \quad \text{for all } d \neq 0. \quad (2.2.5)$$

PROOF. [(i)] Let first  $s \in S$ . We have already seen in Definition 2.1.4 that

$$-\sigma_S(-d) \leq \langle s, d \rangle \leq \sigma_S(d) \quad \text{for all } d \in \mathbb{R}^n.$$



**Fig. 2.2.2.** Affine hulls and orthogonal spaces

If the breadth of  $S$  along  $d$  is zero, we obtain a pair of equalities: for such  $d$ , there holds

$$\langle s, d \rangle = \sigma_S(d),$$

an equality which extends by affine combination to any element  $s \in \text{aff } S$ .

Conversely, let  $s$  satisfy (2.2.3). A first case is when the only  $d$  described in (2.2.3) is  $d = 0$ ; as a consequence of our observations in Definition 2.1.4, there is no affine hyperplane containing  $S$ , i.e.  $\text{aff } S = \mathbb{R}^n$  and there is nothing to prove. Otherwise, there does exist a hyperplane  $H$  containing  $S$ ; it is defined by

$$H := \{p \in \mathbb{R}^n : \langle p, d_H \rangle = \sigma_S(d_H)\}, \quad (2.2.6)$$

for some  $d_H \neq 0$ . We proceed to prove  $\langle s, \cdot \rangle \leq \sigma_H$ .

In fact, the breadth of  $S$  along  $d_H$  is certainly 0, hence  $\langle s, d_H \rangle = \sigma_S(d_H)$  because of (2.2.3), while (2.2.6) shows that  $\sigma_S(d_H) = \sigma_H(d_H)$ . On the other hand, it is obvious that  $\sigma_H(d) = +\infty$  if  $d$  is not collinear to  $d_H$ . In summary, we have proved  $\langle s, d \rangle \leq \sigma_H(d)$  for all  $d$ , i.e.  $s \in H$ . We conclude that our  $s$  is in any affine manifold containing  $S$ :  $s \in \text{aff } S$ .

[*(iii)*] In view of positive homogeneity, we can normalize  $d$  in (2.2.5); accordingly, denote by  $\tilde{B}$  the unit sphere. For  $s \in \text{int } S$ , there exists  $\varepsilon > 0$  such that  $s + \varepsilon d \in S$  for all  $d \in \tilde{B}$ . Then, from the very definition (2.1.1),

$$\sigma_S(d) \geq \langle s + \varepsilon d, d \rangle = \langle s, d \rangle + \varepsilon \quad \text{for all } d \in \tilde{B}.$$

Conversely, let  $s \in \mathbb{R}^n$  be such that

$$\sigma_S(d) - \langle s, d \rangle > 0 \quad \text{for all } d \in \tilde{B}$$

which implies, because  $\sigma_S$  is closed and the unit sphere is compact:

$$0 < \varepsilon := \inf \{\sigma_S(d) - \langle s, d \rangle : d \in \tilde{B}\} \leq +\infty.$$

Thus

$$\langle s, d \rangle + \varepsilon \leq \sigma_S(d) \quad \text{for all } d \in \tilde{B}.$$

Now take  $u$  with  $\|u\| < \varepsilon$ . From the Cauchy-Schwarz inequality, we have for all  $d \in \tilde{B}$

$$\langle s + u, d \rangle = \langle s, d \rangle + \langle u, d \rangle \leq \langle s, d \rangle + \varepsilon \leq \sigma_S(d)$$

and this implies  $s + u \in S$  because of Theorem 2.2.2:  $s \in \text{int } S$  and (iii) is proved.

[*(ii)*] Look at Fig. 2.2.2 again: decompose  $\mathbb{R}^n = V \oplus U$ , where  $V$  is the subspace parallel to  $\text{aff } S$  and  $U = V^\perp$ . In the decomposition  $d = d_V + d_U$ ,  $\langle \cdot, d_U \rangle$  is constant over  $S$ , so  $S$  has 0-breadth along  $d_U$  and

$$\sigma_S(d) = \sup_{s \in S} \langle s, d_V + d_U \rangle = \langle s, d_U \rangle + \sup_{s \in S} \langle s, d_V \rangle$$

for any  $s \in S$ . With these notations, a direction described as in (2.2.4) is a  $d$  such that

$$\sigma_S(d) + \sigma_S(-d) = \sigma_S(d_V) + \sigma_S(-d_V) > 0.$$

Then, (ii) is just (iii) written in the subspace  $V$ .  $\square$

We already know that the effective domain of  $\sigma_S$  is a convex cone, which consists of all oriented directions “in which  $S$  is bounded” (remember Interpretation 2.1.5). This can be made more explicit.

**Proposition 2.2.4** *Let  $S$  be a nonempty closed convex set in  $\mathbb{R}^n$ . Then  $\text{cl dom } \sigma_S$  and the asymptotic cone  $S_\infty$  of  $S$  are mutually polar cones.*

PROOF. Recall from §III.3.2 that, if  $K_1$  and  $K_2$  are two closed convex cones, then  $K_1 \subset K_2$  if and only if  $(K_1)^\circ \supset (K_2)^\circ$ .

Let  $p \in S_\infty$ . Fix  $s_0$  arbitrary in  $S$  and use the fact that  $S_\infty = \cap_{t>0} t(S - s_0)$  (§III.2.2): for all  $t > 0$ , we can find  $s_t \in S$  such that  $p = t(s_t - s_0)$ . Now, for  $q \in \text{dom } \sigma_S$ , there holds

$$\langle p, q \rangle = t \langle s_t - s_0, q \rangle \leq t [\sigma_S(q) - \langle s_0, q \rangle] < +\infty$$

and letting  $t \downarrow 0$  shows that  $\langle p, q \rangle \leq 0$ . In other words,  $\text{dom } \sigma_S \subset (S_\infty)^\circ$ ; then  $\text{cl dom } \sigma_S \subset (S_\infty)^\circ$  since the latter is closed.

Conversely, let  $q \in (\text{dom } \sigma_S)^\circ$ , which is a cone, hence  $tq \in (\text{dom } \sigma_S)^\circ$  for any  $t > 0$ . Thus, given  $s_0 \in S$ , we have for arbitrary  $p \in \text{dom } \sigma_S$

$$\langle s_0 + tq, p \rangle = \langle s_0, p \rangle + t \langle q, p \rangle \leq \langle s_0, p \rangle \leq \sigma_S(p),$$

so  $s_0 + tq \in S$  by virtue of Theorem 2.2.2. In other words

$$q \in \frac{S - s_0}{t} \quad \text{for all } t > 0$$

and  $q \in S_\infty$ .  $\square$

### 2.3 Examples

Let us start with elementary situations. The simplest example of a support function is that of a singleton  $\{s\}$ . Then  $\sigma_{\{s\}}$  is merely  $\langle s, \cdot \rangle$ , we have a first illustration of the introduction (iii) to this chapter: the concept of a linear form  $\langle s, \cdot \rangle$  can be generalized to  $s$  not being a singleton, which amounts to generalizing linearity to closed sublinearity (more details will be given in §3). The case when  $S$  is the unit ball  $B(0, 1)$  is also rather simple:

$$\sigma_{B(0,1)}(d) \geq \left\langle \frac{d}{\|d\|}, d \right\rangle = \|d\| \quad (\text{if } d \neq 0)$$

and, for  $s \in B(0, 1)$ , the Cauchy-Schwarz inequality implies  $\langle s, d \rangle \leq \|d\|$ . Altogether,

$$\sigma_{B(0,1)}(d) = \|d\|. \quad (2.3.1)$$

Our next example is the simplest possible illustration of Proposition 2.2.4, namely when  $S_\infty$  is  $S$  itself.

**Example 2.3.1 (Cones, Half-Spaces, Subspaces)** Let  $K$  be a closed convex cone of  $\mathbb{R}^n$ . Then

$$\sigma_K(d) = \begin{cases} 0 & \text{if } \langle s, d \rangle \leq 0 \text{ for all } s \in K, \\ +\infty & \text{otherwise.} \end{cases}$$

In other words,  $\sigma_K$  is the indicator function of the polar cone  $K^\circ$ . Note the symmetry: since  $K^{\circ\circ} = K$ , the support function of  $K^\circ$  is the indicator of  $K$ .

Two particular cases are of interest. One is when  $K$  is a half-space:

$$K := \{s \in \mathbb{R}^n : \langle s, v \rangle \leq 0\};$$

then it is clear enough that

$$\sigma_K(d) = \begin{cases} 0 & \text{if } d = tv \text{ with } t \geq 0, \\ +\infty & \text{otherwise.} \end{cases} \quad (2.3.2)$$

Needless to say, the support function of the half-line  $\mathbb{R}^+v$  (the polar of  $K$ ) is in turn the indicator of  $K$ .

The other interesting case is that of a subspace. Let  $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a linear operator and  $H$  be defined by

$$H := \text{Ker } A = \{s \in \mathbb{R}^n : As = 0\}.$$

Then the support function of  $H$  is the indicator of the orthogonal subspace  $H^\perp$ :

$$\sigma_H(d) = I_{H^\perp}(d) = \begin{cases} 0 & \text{if } \langle s, d \rangle = 0 \text{ for all } s \in H, \\ +\infty & \text{otherwise.} \end{cases}$$

The subspace  $H^\perp$  can be defined with the help of the adjoint of  $A$ :

$$H^\perp = (\text{Ker } A)^\perp = \text{Im } A^* = \{A^*\lambda : \lambda \in \mathbb{R}^m\}.$$

If  $A$  or  $H$  are defined in terms of linear constraints

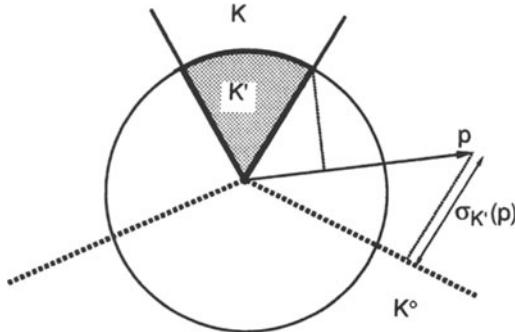
$$H := \{s \in \mathbb{R}^n : \langle s, a_j \rangle = 0 \text{ for } j = 1, \dots, m\},$$

then

$$H^\perp = \left\{ \sum_{j=1}^m \lambda_j a_j : \lambda \in \mathbb{R}^m \right\}.$$

All these calculations are useful in constrained optimization, where one often deals with closed convex polyhedra expressed as intersections of half-spaces and subspaces.

Figure 2.3.1 illustrates a modification in which our cone  $K$  is modified to  $K' := K \cap B(0, 1)$ . The calculus rules of §3.3 will prove what is suggested by the picture: the support function of  $K'$  is the distance function to  $K^\circ$  (check the similarity of the appropriate triangles, and note that  $\sigma_{K'}(d) = 0$  when  $d \in K^\circ$ ).  $\square$



**Fig. 2.3.1.** Support function of a truncated cone

### Example 2.3.2 Set

$$S := \{s = (\rho, \tau) \in \mathbb{R}^2 : \rho > 0, \tau \geq 1/\rho\}. \quad (2.3.3)$$

Its asymptotic cone is

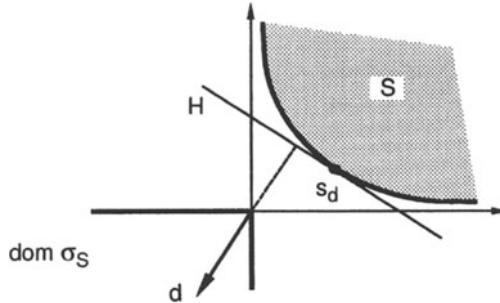
$$S_\infty = \{(\rho, \tau) \in \mathbb{R}^2 : \rho \geq 0, \tau \geq 0\}$$

and, from Proposition 2.2.4:

$$\text{dom } \sigma_S \subset \{(\xi, \eta) : \xi \leq 0, \eta \leq 0\}.$$

The exact status of the boundary of  $\text{dom } \sigma_S$  (i.e. when  $\xi\eta = 0$ ) is not specified by Proposition 2.2.4: is  $\sigma_S$  finite there? The computation of  $\sigma_S$  can be done directly from the definitions (2.1.1) and (2.3.3). The following geometrical argument yields simpler calculations, however (see Fig. 2.3.2): for given  $d = (\xi, \eta) \neq (0, 0)$ , consider the hyperplane

$$H_{d, \sigma_S(d)} = \{(\alpha, \beta) : \xi\alpha + \eta\beta = \sigma_S(d)\}.$$

**Fig. 2.3.2.** A support function

It has to be tangent to the boundary of  $S$ , defined by the equation  $\alpha\beta = 1$ . So, the discriminant  $\sigma_S^2(d) - 4\xi\eta$  of the equation in  $\alpha$

$$\xi\alpha + \eta\frac{1}{\alpha} = \sigma_S(d)$$

must be 0. We obtain directly  $\sigma_S(\xi, \eta) = -2\sqrt{\xi\eta}$  for  $\xi < 0, \eta < 0$  (the sign is “-” because  $0 \notin S$ ; remember Theorem 2.2.2). Finally, Proposition 2.1.2 tells us that the closed function  $\sigma_S(\xi, \eta)$  has to be 0 when  $\xi\eta = 0$ . All this is confirmed by Fig. 2.3.2.  $\square$

**Remark 2.3.3** Two features concerning the boundary of  $\text{dom } \sigma_S$  are worth mentioning on the above example: the supremum in (2.1.1) is not attained when  $d \in \text{bd dom } \sigma_S$  (the point  $s_d$  of Fig. 2.1.1 is sent to infinity when  $d$  approaches  $\text{bd dom } \sigma_S$ ), and  $\text{dom } \sigma_S$  is closed.

These are not the only possible cases: Example 2.3.1 shows that the supremum in (2.1.1) can well be attained for all  $d \in \text{dom } \sigma_S$ ; and in the example

$$S := \{(\rho, \tau) : \tau \geq \frac{1}{2}\rho^2\},$$

$\text{dom } \sigma_S$  is not closed. The difference is that, now,  $S$  has no asymptote “at finite distance”.  $\square$

**Example 2.3.4** (cf. Example 1.2.3) Let  $Q$  be a symmetric positive definite operator from  $\mathbb{R}^n$  to  $\mathbb{R}^n$  and consider the sublevel-set

$$E_Q := \{s \in \mathbb{R}^n : \langle Qs, s \rangle \leq 1\}.$$

The support function of  $E_Q$  is defined by

$$d \mapsto \sigma_{E_Q}(d) := \max \{\langle s, d \rangle : \langle Qs, s \rangle \leq 1\}. \quad (2.3.4)$$

Calling  $Q^{1/2}$  the square root of  $Q$ , the change of variable  $p = Q^{1/2}s$  in (2.3.4) gives

$$\sigma_{E_Q}(d) = \max \left\{ \langle p, Q^{-1/2}d \rangle : \|p\|^2 \leq 1 \right\}$$

whose unique solution for  $d \neq 0$  (again Cauchy-Schwarz!) is  $p = \frac{Q^{-1/2}d}{\|Q^{-1/2}d\|}$  and finally

$$\sigma_{E_Q}(d) = \|Q^{-1/2}d\| = \sqrt{\langle d, Q^{-1}d \rangle}. \quad (2.3.5)$$

Observe in this example the “duality” between the gauge  $x \mapsto \sqrt{\langle Qx, x \rangle}$  of  $E_Q$  and its support function (2.3.5).

When  $Q$  is merely symmetric positive semi-definite,  $E_Q$  becomes an elliptic cylinder, whose asymptotic cone is  $\text{Ker } Q$  (remember Example 1.2.3). Then Proposition 2.2.4 tells us that

$$\text{cl dom } \sigma_{E_Q} = (\text{Ker } Q)^\circ = (\text{Ker } Q)^\perp = \text{Im } Q.$$

When  $d \in \text{Im } Q$ ,  $\sigma_{E_Q}(d)$  is finite indeed and (2.3.5) does hold,  $Q^{-1}d$  denoting now any element  $p$  such that  $Qp = d$ . We leave this as an exercise.  $\square$

### 3 The Isomorphism Between Closed Convex Sets and Closed Sublinear Functions

#### 3.1 The Fundamental Correspondence

We have seen in Proposition 2.1.2 that a support function is closed and sublinear. What about the converse? Are there closed sublinear functions which support no set in  $\mathbb{R}^n$ ? The answer is no: any closed sublinear function *can be viewed* as a support function. The key lies in the representation of a closed convex function  $f$  via affine functions minorizing it: when the starting  $f$  is also positively homogeneous, the underlying affine functions can be assumed linear.

**Theorem 3.1.1** *Let  $\sigma$  be a closed sublinear function; then there is a linear function minorizing  $\sigma$ . In fact,  $\sigma$  is the supremum of the linear functions minorizing it. In other words,  $\sigma$  is the support function of the nonempty closed convex set*

$$S_\sigma := \{s \in \mathbb{R}^n : \langle s, d \rangle \leq \sigma(d) \text{ for all } d \in \mathbb{R}^n\}. \quad (3.1.1)$$

PROOF. Being convex,  $\sigma$  is minorized by some affine function (Proposition IV.1.2.1): for some  $(s, r) \in \mathbb{R}^n \times \mathbb{R}$ ,

$$\langle s, d \rangle - r \leq \sigma(d) \quad \text{for all } d \in \mathbb{R}^n. \quad (3.1.2)$$

Because  $\sigma(0) = 0$ , the above  $r$  is nonnegative. Also, by positive homogeneity,

$$\langle s, d \rangle - \frac{1}{t}r \leq \sigma(d) \quad \text{for all } d \in \mathbb{R}^n \text{ and all } t > 0.$$

Letting  $t \rightarrow +\infty$ , we see that  $\sigma$  is actually minorized by a linear function:

$$\langle s, d \rangle \leq \sigma(d) \quad \text{for all } d \in \mathbb{R}^n. \quad (3.1.3)$$

Now observe that the minorization (3.1.3) is sharper than (3.1.2): when expressing the closed convex  $\sigma$  as the supremum of all the affine functions minorizing it (Proposition IV.1.2.8), we can restrict ourselves to linear functions. In other words

$$\sigma(d) = \sup \{ \langle s, d \rangle : \text{the linear } \langle s, \cdot \rangle \text{ minorizes } \sigma \};$$

in the above index-set, we just recognize  $S_\sigma$ .  $\square$

One of the important points in this result is the *nonemptiness* of  $S_\sigma$  in (3.1.1); we have here the analytical form of Hahn-Banach theorem: there *exists* a linear function minorizing the closed sublinear function  $\sigma$ .

Another way of expressing Theorem 3.1.1 is that the closed convex set  $\text{epi } \sigma$  is the intersection of the closed half-spaces containing it; but since  $\text{epi } \sigma$  is actually a cone, these half-spaces can be assumed to have *linear* hyperplanes as boundaries (remember Remark III.4.2.8). A connection between  $S_\sigma$  and the cone polar to  $\text{epi } \sigma$  is thus introduced; Chap. VI will exploit this remark.

The main consequence of this important theorem is an assessment of closed sublinear functions. Section 2.2 has established a bijection from closed convex sets onto support functions. Thanks to Theorem 3.1.1, this bijection is actually onto *closed sublinear functions*, which is of course much more satisfactory: the latter class of functions is defined in abstracto, while the former class was ad hoc, as far as this bijection was concerned.

Thus, the wording “support function” in Fig. 2.2.1 can everywhere be replaced by “closed sublinear”. This replacement can be done in Theorem 2.2.2 as well:

**Corollary 3.1.2** *For a nonempty closed convex set  $S$  and a closed sublinear function  $\sigma$ , the following are equivalent:*

- (i)  $\sigma$  is the support function of  $S$ ,
- (ii)  $S = \{s : \langle s, d \rangle \leq \sigma(d) \text{ for all } d \in X\}$ , where the set  $X$  can be indifferently taken as: the whole of  $\mathbb{R}^n$ , the unit ball  $B(0, 1)$  or its boundary, or  $\text{dom } \sigma$ .

PROOF. The case  $X = \mathbb{R}^n$  is just Theorem 3.1.1. The other cases are then clear.  $\square$

Remember §III.4.2(b): a closed convex set  $S$  is geometrically characterized as an intersection of half-spaces, which in turn can be characterized in terms of the support function of  $S$ . Each  $(d, r) \in \mathbb{R}^n \times \mathbb{R}$  defines (for  $d \neq 0$ ) the half-space  $H_{d,r}^-$  via (2.1.3). This half-space contains  $S$  if and only if  $r \geq \sigma(d)$ , and Corollary 3.1.2 expresses that

$$S = \cap \{s : \langle s, d \rangle \leq r \text{ for all } d \in \mathbb{R}^n \text{ and } r \geq \sigma(d)\},$$

in which the couple  $(d, r)$  plays the role of an index, running in the index-set  $\text{epi } \sigma \subset \mathbb{R}^n \times \mathbb{R}$  (compare with the discussion after Definition 2.1.1). Of course, this index-set can be reduced down to  $\mathbb{R}^n$ : the above formula can be simplified to

$$S = \cap \{s : \langle s, d \rangle \leq \sigma(d) \text{ for all } d \in X\}$$

where  $X$  can be taken as in Corollary 3.1.2.

Recall from §III.2.4 that an exposed face of a convex set  $S$  is defined as the set of points of  $S$  which maximize some (nonzero) linear form. This concept appears as particularly welcome in the context of support functions:

**Definition 3.1.3** Let  $S$  be a nonempty closed convex set, with support function  $\sigma$ . For given  $d \neq 0$ , the set

$$F_S(d) := \{s \in S : \langle s, d \rangle = \sigma(d)\}$$

is called the exposed face of  $S$  associated with  $d$ , or the *face exposed by  $d$* .

□

For a unified notation, the entire  $S$  can be considered as the face exposed by 0. On the other hand, a given  $d$  may expose no face at all (when  $S$  is unbounded).

Symmetrically to Definition 3.1.3, one can ask what are those  $d \in \mathbb{R}^n$  such that  $\langle \cdot, d \rangle$  is maximized at a given  $s \in S$ . We obtain nothing other than the normal cone  $N_S(s)$  to  $S$  at  $s$ , as is obvious from its Definition III.5.2.3. The following result is simply a restatement of Proposition III.5.3.3.

**Proposition 3.1.4** For  $s$  in a nonempty closed convex set  $S$ , it holds

$$s \in F_S(d) \iff d \in N_S(s).$$

□

When  $d$  describes the set of normalized directions, the corresponding exposed faces exactly describe the boundary of  $S$ :

**Proposition 3.1.5** For a nonempty closed convex set  $S$ , it holds

$$\text{bd } S = \bigcup \{F_S(d) : d \in X\}$$

where  $X$  can be indifferently taken as:  $\mathbb{R}^n \setminus \{0\}$ , the unit sphere  $\widetilde{B}$ , or  $\text{dom } \sigma_S \setminus \{0\}$ .

PROOF. Observe from Definition 3.1.3 that the face exposed by  $d \neq 0$  does not depend on  $\|d\|$ . This establishes the equivalence between the first two choices for  $X$ . As for the third choice, it is due to the fact that  $F_S(d) = \emptyset$  if  $d \notin \text{dom } \sigma_S$ .

Now, if  $s$  is interior to  $S$  and  $d \neq 0$ , then  $s + \epsilon d \in S$  and  $s$  cannot be a maximizer of  $\langle \cdot, d \rangle$ :  $s$  is not in the face exposed by  $d$ . Conversely, take  $s$  on the boundary of  $S$ . Then  $N_S(s)$  contains a nonzero vector  $d$ ; by Proposition 3.1.4,  $s \in F_S(d)$ .

□

### 3.2 Example: Norms and Their Duals, Polarity

Let  $\|\cdot\|$  be an arbitrary norm on  $\mathbb{R}^n$ . It is a positive (except at 0) closed sublinear function and its sublevel-set

$$B := \{x \in \mathbb{R}^n : \|x\| \leq 1\} \tag{3.2.1}$$

is particularly interesting. It is the unit ball associated with the norm, a symmetric, convex, compact set containing the origin as an interior point;  $\|\cdot\|$  is the gauge of  $B$  (§1.2). On the other hand, why not take the set whose support function is  $\|\cdot\|$ ? In view of Corollary 3.1.2, it is defined by

$$\{s \in \mathbb{R}^n : \langle s, x \rangle \leq \|x\| \text{ for all } x \in \mathbb{R}^n\} =: B^*. \tag{3.2.2}$$

It is an easy exercise to check that  $B^*$  is also symmetric, convex, compact; and it contains the origin as an interior point (Theorem 2.2.3(iii)).

Now, we have two closed convex sets  $B$  and  $B^*$ . We can generate two more closed sublinear functions: take the support function  $\sigma_B$  of  $B$  and the gauge  $\gamma_{B^*}$  of  $B^*$ . It turns out that we then obtain the same function, which actually is a norm, denoted  $\|\cdot\|^*$ : the so-called *dual norm* of  $\|\cdot\|$ . The game finishes there: the two sets that  $\|\cdot\|^*$  supports and is the gauge of, respectively, are  $B$  and  $B^*$ .

**Proposition 3.2.1** *Let  $B$  and  $B^*$  be defined by (3.2.1) and (3.2.2), where  $\|\cdot\|$  is a norm on  $\mathbb{R}^n$ . The support function of  $B$  and the gauge of  $B^*$  are the same function  $\|\cdot\|^*$  defined by*

$$\|s\|^* := \max \{\langle s, x \rangle : \|x\| \leq 1\}. \quad (3.2.3)$$

Furthermore,  $\|\cdot\|^*$  is a norm on  $\mathbb{R}^n$ . The support function of its unit ball  $B^*$  and the gauge of its supported set  $B$  are the same function  $\|\cdot\|$ : there holds

$$\|x\| = \max \{\langle s, x \rangle : \|s\|^* \leq 1\}. \quad (3.2.4)$$

PROOF. It is a particular case of the results 3.2.4 and 3.2.5 below.  $\square$

Note the following symmetric relation (“Cauchy-Schwarz”)

$$\langle s, x \rangle \leq \|s\|^* \|x\| \quad \text{for all } (s, x) \in \mathbb{R}^n \times \mathbb{R}^n, \quad (3.2.5)$$

which comes directly from (3.2.3), using positive homogeneity. It expresses the duality correspondence between the two Banach spaces  $(\mathbb{R}^n, \|\cdot\|)$  and  $(\mathbb{R}^n, \|\cdot\|^*)$ . Furthermore, equality holds in (3.2.5) when  $s \neq 0$  and  $x \neq 0$  form an associated pair via Proposition 3.1.4:

$$\frac{s}{\|s\|^*} \in F_{B^*}(x) \quad \text{or equivalently} \quad \frac{x}{\|x\|} \in F_B(s).$$

Thus, a norm automatically defines another norm (its dual); and the operation is symmetric: the dual of the dual norm is the norm itself.

**Remark 3.2.2** The operation (3.2.3)–(3.2.4) establishes a “duality” correspondence within a subclass of closed sublinear functions: those that are symmetric, finite everywhere, and positive (except at 0) – in short, norms.

This analytic operation has its counterpart in the geometric world: starting from a closed convex set which is symmetric, bounded and contains the origin as an interior point – in a word, a “unit ball” – such as  $B$ , one constructs via gauges and support functions another closed convex set  $B^*$  which has the same properties. This correspondence is called *polarity*, demonstrated by Fig. 3.2.1: the polar (set) of  $B$  is

$$B^* := \{s : \langle s, x \rangle \leq 1 \text{ for all } x \in B\} \quad (3.2.6)$$

and symmetrically, the polar of  $B^*$  is

$$(B^*)^* := \{x : \langle s, x \rangle \leq 1 \text{ for all } s \in B^*\} = B. \quad (3.2.7)$$

$\square$

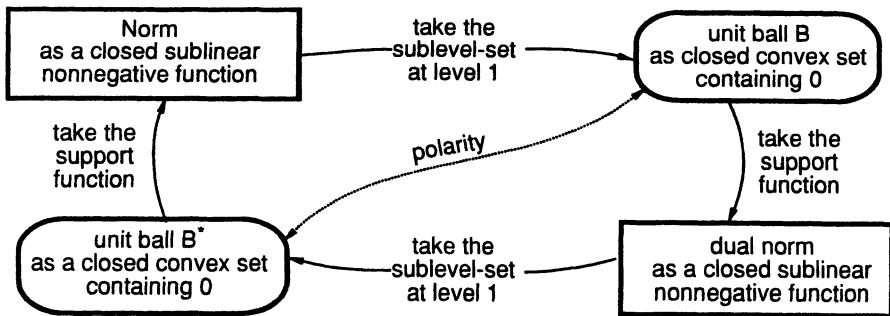


Fig. 3.2.1. Dual norms and polar sets

We leave it as an exercise to draw the unit balls of the  $\ell_1$ - and  $\ell_\infty$ -norms on  $\mathbb{R}^n$ :

$$\|x\|_1 := \sum_{i=1}^n |x^i| \quad \text{and} \quad \|x\|_\infty := \max \{|x^1|, \dots, |x^n|\}$$

(proceed as in Interpretation 2.1.5: a picture in  $\mathbb{R}^n$  will do). Observe on the picture thus obtained that they are in polarity correspondence if the scalar product is the usual dot-product  $\langle x, y \rangle = x^\top y$ .

A more complicated situation is illustrated by the “hexagonal norm” of Fig. 3.2.2. Observe how elongation in one direction corresponds to contraction for the polar. Also: a facet of one of the sets is exposed by a vertex in the polar.

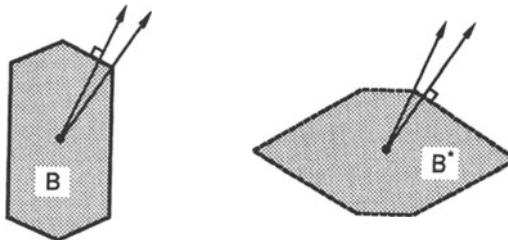


Fig. 3.2.2. Hexagonal unit-balls

**Example 3.2.3** Other important norms are the quadratic norms, defined by

$$\|x\|_Q := \sqrt{\langle Qx, x \rangle}$$

where  $Q$  is a symmetric positive definite linear operator. They are important because they derive from a scalar product on  $\mathbb{R}^n$ , namely:

$$\langle x, y \rangle_Q := \langle Qx, y \rangle.$$

We refer to Example 2.3.4, more precisely formula (2.3.5), to compute the corresponding dual norm

$$(\|s\|_Q)^* = \sqrt{\langle s, Q^{-1}s \rangle} = \|s\|_{Q^{-1}}.$$

When  $Q = I_n$ , we get back the Euclidean norm  $\langle \cdot, \cdot \rangle^{1/2}$ . A comparison of (2.3.1) and (3.2.3) shows that it is self-dual:  $\|\cdot\|^* = \|\cdot\|$ . Among all the possible norms on  $\mathbb{R}^n$ , it is the only one having this property (once the scalar product is chosen!).  $\square$

Actually, polarity neither relies upon symmetry, nor boundedness, nor on having 0 as an interior point. To take gauges and support functions resulting in (3.2.6) – (3.2.7), the only important property is after all that 0 be in the closed convex set under consideration ( $B$  or  $B^*$ ). In other words, the polarity relations (3.2.6), (3.2.7) establish an involution between sets that are merely closed convex, and contain the origin. More precisely, we have the following result:

**Proposition 3.2.4** *Let  $C$  be a closed convex set containing the origin. Its gauge  $\gamma_C$  is the support function of a closed convex set containing the origin, namely*

$$C^\circ := \{s \in \mathbb{R}^n : \langle s, d \rangle \leq 1 \text{ for all } d \in C\}, \quad (3.2.8)$$

which defines the polar (set) of  $C$ .

PROOF. We know that  $\gamma_C$  (which, by Theorem 1.2.5(i), is closed, sublinear and non-negative) is the support function of some closed convex set containing the origin, say  $D$ ; from (3.1.1),

$$D = \{s \in \mathbb{R}^n : \langle s, d \rangle \leq r \text{ for all } (d, r) \in \text{epi } \gamma_C\}.$$

As seen in (1.2.4),  $\text{epi } \gamma_C$  is the closed convex conical hull of  $C \times \{1\}$ ; we can use positive homogeneity to write

$$D = \{s \in \mathbb{R}^n : \langle s, d \rangle \leq 1 \text{ for all } d \text{ such that } \gamma_C(d) \leq 1\}.$$

In view of Theorem 1.2.5(iii), the above index-set is just  $C$ ; in other words,  $D = C^\circ$ .  $\square$

Geometrically, the above proof is illustrated by Fig. 3.2.3, in which dual elements are drawn in dashed lines:  $D = C^\circ$  is obtained by cutting the polar cone  $(\text{epi } \gamma_C)^\circ$  at the level  $-1$ . Turn the picture upside down: cutting the polar cone  $(\text{epi } \gamma_{C^\circ})^\circ$  at the level which has now become  $-1$ , we obtain  $(C^\circ)^\circ$ . But the polarity between closed convex cones is involutive: the picture shows that  $(\text{epi } \gamma_{C^\circ})^\circ$  is our original cone  $\text{epi } \gamma_C$ . In other words,  $C^{\circ\circ} = C$ , Proposition 3.2.4 has its dual version:

**Corollary 3.2.5** *Let  $C$  be a closed convex set containing the origin. Its support function  $\sigma_C$  is the gauge of  $C^\circ$ .*  $\square$

**Remark 3.2.6** The elementary operation making up polarity is a one-to-one mapping between nonzero vectors and affine hyperplanes not containing the origin, via the equation inspired from (3.2.8):

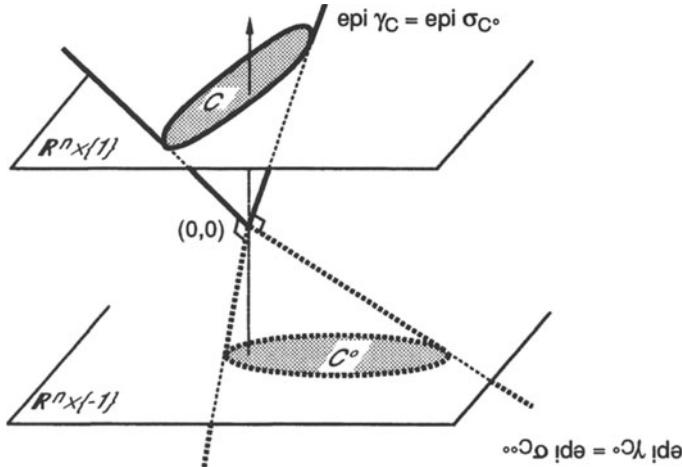


Fig. 3.2.3. Gauges and supports

$$s \mapsto H(s) := H_{s,1} = \{y \in \mathbb{R}^n : \langle s, y \rangle = 1\}. \quad (3.2.9)$$

Direct calculations show for example that the polar of the half-space

$$H^- := \{y = (\xi, \eta) \in \mathbb{R}^2 : \xi \leq 2\}$$

is the segment

$$(H^-)^\circ = \{(\rho, 0) : 0 \leq \rho \leq 1/2\}.$$

This simple example suggests the following comment: if \$\sigma\$ is a given nonnegative closed sublinear function, it is the gauge of a set \$G\$ which can be immediately constructed: along \$0 \neq s \in \mathbb{R}^n\$, plot the point \$g(s) = s/\sigma(s) \in [0, +\infty]s\$. Then \$G\$ is the union of the segments \$[0, g(s)]\$, with \$s\$ describing the unit sphere. If, along the same \$s\$, we plot the point \$\sigma(s)s\$, we likewise get a description of the set \$S\$ supported by \$\sigma\$, but in a much less direct way: \$G\$ is now *enveloped* by the affine hyperplane orthogonal to \$s\$ and containing the point \$\sigma(s)s\$; now, *differentiation* is involved.

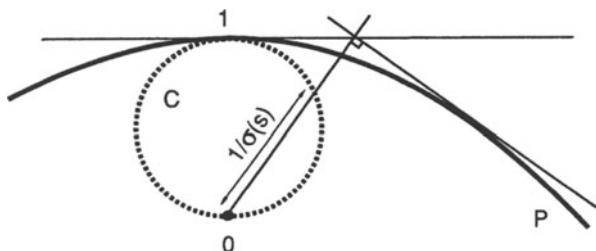


Fig. 3.2.4. Description of mutually polar sets

An expert in geometry will for example see on Fig. 3.2.4 that the polar of the circle

$$C = \{(\rho, \tau) : \rho^2 + (\tau - 1/2)^2 \leq 1/4\}$$

has a parabolic boundary. We leave it as an exercise to compute the gauge of  $C$ , and to realize that it is the support function of

$$P = \{(\xi, \eta) : \xi^2 \leq 1 - \eta\}.$$

Constructing a set from its gauge thus appears to be substantially easier than it is from its support function. Furthermore, to make a support function, we need a scalar product, while a gauge just needs an origin in  $\mathbb{R}^n$ . These advantages, however, are balanced by the rich calculus which can be developed with support functions, and which will be the subject of §3.3.  $\square$

It is clear from (3.2.8) that, for all  $(d, s) \in C \times C^\circ$ ,  $\langle s, d \rangle \leq 1$ ; this implies in particular that no nonzero  $s \in C^\circ$  can be in the asymptotic cone of  $C$ . Furthermore, the property  $\langle s, d \rangle = 1$  means that  $d$  exposes in  $C^\circ$  a face  $F_{C^\circ}(d)$  containing  $s$ ; and  $s$  exposes likewise in  $C^{oo} = C$  a face  $F_C(s)$  containing  $d$ . Because the boundary of a closed convex set is described by its exposed faces (Proposition 3.1.5), the following result is then natural; compare it with Fig. 3.2.2.

**Proposition 3.2.7** *Let  $C$  be a nonempty compact convex set having 0 in its interior, so that  $C^\circ$  enjoys the same properties. Then, for all  $d$  and  $s$  in  $\mathbb{R}^n$ , the following statements are equivalent (the notation (3.2.9) is used)*

- (i)  $H(s)$  is a supporting hyperplane to  $C$  at  $d$ ;
- (ii)  $H(d)$  is a supporting hyperplane to  $C^\circ$  at  $s$ ;
- (iii)  $d \in \text{bd } C$ ,  $s \in \text{bd } C^\circ$  and  $\langle s, d \rangle = 1$ ;
- (iv)  $d \in C$ ,  $s \in C^\circ$  and  $\langle s, d \rangle = 1$ .

PROOF. Left as an exercise; the assumptions are present to make sure that every nonzero vector in  $\mathbb{R}^n$  does expose a face in each set.  $\square$

Finally, suppose that  $C$  in (3.2.8) is a cone. By positive homogeneity, the number “1” can be replaced by any positive number, and even by “0” (remember the proof of Theorem 3.1.1). We recognize the definition of polarity between closed convex cones.

### 3.3 Calculus with Support Functions

From §1.3, the set of sublinear functions has a structure allowing calculus. Likewise, a calculus exists with subsets of  $\mathbb{R}^n$ . Then a natural question is: to what extent are these structures in correspondence via the supporting operation? In other words, to what extent is the supporting operation an isomorphism? The answer turns out to be very rich indeed.

We start with the order relation

**Theorem 3.3.1** *Let  $S_1$  and  $S_2$  be nonempty closed convex sets; call  $\sigma_1$  and  $\sigma_2$  their support functions. Then*

$$S_1 \subset S_2 \iff \sigma_1(d) \leq \sigma_2(d) \text{ for all } d \in \mathbb{R}^n.$$

PROOF. Apply the equivalence stated in Corollary 3.1.2:

$$\begin{aligned} S_1 \subset S_2 &\iff s \in S_2 \text{ for all } s \in S_1 \\ &\iff \sigma_2(d) \geq \langle s, d \rangle \text{ for all } s \in S_1 \text{ and all } d \in \mathbb{R}^n \\ &\iff \sigma_2(d) \geq \sup_{s \in S_1} \langle s, d \rangle \text{ for all } d \in \mathbb{R}^n. \end{aligned}$$

□

In a way, the above result generalizes Theorem 2.2.2. It can be supplemented with a partial ordering rule:

**Corollary 3.3.2** *Let  $P_V(S)$  denote the projection of a set  $S$  onto a fixed subspace  $V$ . If  $S_1$  and  $S_2$  are nonempty closed and convex,*

$$\text{cl}(P_V(S_1)) \subset \text{cl}(P_V(S_2)) \iff \sigma_{S_1} \leq \sigma_{S_2} \text{ on } V. \quad (3.3.1)$$

PROOF. Theorem 3.3.1 tells us that the first inclusion is equivalent to

$$\sigma_{P_V(S_1)}(d) \leq \sigma_{P_V(S_2)}(d) \text{ for all } d \in \mathbb{R}^n,$$

which actually means

$$\sigma_{P_V(S_1)}(d) \leq \sigma_{P_V(S_2)}(d) \text{ for all } d \in V. \quad (3.3.2)$$

The reason is that, in terms of the decomposition  $d = d_V + d_{V^\perp}$ , we have  $\langle s, d \rangle = \langle s, d_V \rangle$  for all  $s \in V$ . This equality is transmitted to the supremum over  $S_1$  or  $S_2$ , both sets being included in  $V$ .

Now write any  $s_V \in P_V(S_i)$  as  $s_V = s - s_{V^\perp}$  with  $s \in S_i$ :

$$\sigma_{P_V(S_i)}(d) = \sigma_{S_i}(d) \text{ for } i = 1, 2 \text{ and } d \in V.$$

Using these equalities in (3.3.2), the equivalence (3.3.1) is established. □

The next statement goes with Propositions 1.3.1 and 1.3.2.

### Theorem 3.3.3

(i) *Let  $\sigma_1$  and  $\sigma_2$  be the support functions of the nonempty closed convex sets  $S_1$  and  $S_2$ . If  $t_1$  and  $t_2$  are positive, then*

*$t_1\sigma_1 + t_2\sigma_2$  is the support function of  $\text{cl}(t_1S_1 + t_2S_2)$ .*

(ii) *Let  $\{\sigma_j\}_{j \in J}$  be the support functions of the family of nonempty closed convex sets  $\{S_j\}_{j \in J}$ . Then*

*$\sup_{j \in J} \sigma_j$  is the support function of  $\overline{\text{co}} \{ \cup S_j : j \in J \}$ .*

(iii) *Let  $\{\sigma_j\}_{j \in J}$  be the support functions of the family of closed convex sets  $\{S_j\}_{j \in J}$ . If*

$$S := \bigcap_{j \in J} S_j \neq \emptyset,$$

*then*

$$\sigma_S = \overline{\text{co}} \{ \inf \sigma_j : j \in J \}.$$

PROOF. [(i)] Call  $S$  the closed convex set  $\text{cl}(t_1 S_1 + t_2 S_2)$ . By definition, its support function is

$$\sigma_S(d) = \sup \{\langle t_1 s_1 + t_2 s_2, d \rangle : s_1 \in S_1, s_2 \in S_2\}.$$

In the above expression,  $s_1$  and  $s_2$  run independently in their index sets  $S_1$  and  $S_2$ ,  $t_1$  and  $t_2$  are positive, so

$$\sigma_S(d) = t_1 \sup_{s \in S_1} \langle s, d \rangle + t_2 \sup_{s \in S_2} \langle s, d \rangle.$$

[(ii)] The support function of  $S := \bigcup_{j \in J} S_j$  is

$$\sup_{s \in \bigcup S_j} \langle s, d \rangle = \sup_{j \in J} [\sup_{s_j \in S_j} \langle s_j, d \rangle] = \sup_{j \in J} \sigma_j(d).$$

This implies (ii) since  $\sigma_S = \sigma_{\text{co}} S$ .

[(iii)] The set  $S := \cap S_j$  being nonempty, it has a support function  $\sigma_S$ . Now, from Corollary 3.1.2,

$$\begin{aligned} s \in S &\iff s \in S_j \text{ for all } j \in J \\ &\iff \langle s, \cdot \rangle \leq \sigma_j \text{ for all } j \in J \\ &\iff \langle s, \cdot \rangle \leq \inf_{j \in J} \sigma_j \iff \langle s, \cdot \rangle \leq \text{co}(\inf_{j \in J} \sigma_j) \end{aligned}$$

where the last equivalence comes directly from the Definition IV.2.5.3 of a closed convex hull. Again Corollary 3.1.2 tells us that the closed sublinear function  $\text{co}(\inf_{j \in J} \sigma_j)$  is just the support function of  $S$ .  $\square$

- It is important to observe in (i) that, if  $S_2$  is bounded, then  $t_1 S_1 + t_2 S_2$  is automatically closed. This addition rule can be used to complete Corollary 3.3.2: the right-hand side in (3.3.1) exactly means

$$\sigma_{S_1} + I_V \leq \sigma_{S_2} + I_V;$$

now use Example 2.3.1:  $I_V = \sigma_{V^\perp}$  and finally, (3.3.1) is further equivalent to

$$\text{cl}(S_1 + V^\perp) \subset \text{cl}(S_2 + V^\perp). \quad (3.3.3)$$

- As for (iii), we have seen in Proposition 1.3.2(ii) that, if  $J = \{1, \dots, m\}$  is a finite set, then the “co” operation can be replaced by the infimal convolution: there holds

$$\sigma_{S_1 \cap \dots \cap S_m} = \text{cl}(\sigma_1 \downarrow \dots \downarrow \sigma_m). \quad (3.3.4)$$

This last formula is a simplification of (iii), but the closure operation should not be forgotten, and it is something really complicated; these issues will be addressed more thoroughly in §X.2.3.

- Let  $K$  be a closed convex cone and, as in the end of Example 2.3.1, take  $K' := K \cap B(0, 1)$ . In view of the above observation, the support function of  $K'$  is given by an inf-convolution:

$$\sigma_{K'}(d) = \text{cl}\{\inf_y [\sigma_K(y) + \sigma_B(d - y)]\}.$$

Since  $\sigma_K = I_{K^\circ}$ , the infimum forces  $y$  to be in  $K^\circ$ , in which case  $\sigma_K$  vanishes; knowing that  $\sigma_{B(0,1)} = \|\cdot\|$ , the infimum is

$$\inf \{\|d - y\| : y \in K^\circ\}.$$

Here, we are in a favourable case: this infimum is actually a minimum – achieved at the projection  $p_{K^\circ}(d)$  – and the result is a finite convex function, hence continuous; the closure operation is useless and can be omitted. In a word,

$$\sigma_{K \cap B(0,1)} = d_{K^\circ}. \quad (3.3.5)$$

- Positive homogeneity can also be exploited in Theorem 3.3.3(i) to write

$$\sigma_{tS}(d) = \sigma_S(td) \quad \text{for all } d \in \mathbb{R}^n \text{ and } t > 0,$$

a formula which also holds for negative  $t$  (just write the definition). More generally:

**Proposition 3.3.4** *Let  $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$  be a linear operator,  $\mathbb{R}^m$  being equipped with a scalar product  $\langle\langle \cdot, \cdot \rangle\rangle$  for which  $A^*$  is the adjoint of  $A$ . For  $S \subset \mathbb{R}^n$  nonempty, we have*

$$\sigma_{\text{cl } A(S)}(y) = \sigma_S(A^*y) \quad \text{for all } y \in \mathbb{R}^m.$$

PROOF. Just write the definitions

$$\sigma_{A(S)}(y) = \sup_{s \in S} \langle\langle As, y \rangle\rangle = \sup_{s \in S} \langle s, A^*y \rangle$$

and use Proposition 2.2.1 to obtain the result.  $\square$

Taking an image-function (see §IV.2.4) is another operation involving a linear operator. Its status is slightly more delicate.

**Proposition 3.3.5** *Let  $A : \mathbb{R}^m \rightarrow \mathbb{R}^n$  be a linear operator,  $\mathbb{R}^m$  being equipped with a scalar product  $\langle\langle \cdot, \cdot \rangle\rangle$  for which  $A^*$  is the adjoint of  $A$ . Let  $\sigma$  be the support function of a nonempty closed convex set  $S \subset \mathbb{R}^m$ . If  $\sigma$  is minorized on the inverse image*

$$\bar{A}(d) = \{p \in \mathbb{R}^m : Ap = d\} \quad (3.3.6)$$

*of each  $d \in \mathbb{R}^n$ , then the support function of the set  $(\bar{A}^*)(S)$  is the closure of the image-function  $A\sigma$ .*

PROOF. Our assumption is tailored to make sure that  $A\sigma \in \text{Conv } \mathbb{R}^n$  (see Theorem IV.2.4.2). The positive homogeneity of  $A\sigma$  is clear: for  $d \in \mathbb{R}^n$  and  $t > 0$ ,

$$(A\sigma)(td) = \inf_{Ap=td} \sigma(p) = \inf_{A(p/t)=d} t\sigma(p/t) = t \inf_{Aq=d} \sigma(q) = t(A\sigma)(d).$$

Thus, the closed sublinear function  $\text{cl}(A\sigma)$  supports some set  $S'$ ; by definition,  $s \in S'$  if and only if

$$\langle s, d \rangle \leq \inf \{\sigma(p) : Ap = d\} \quad \text{for all } d \in \mathbb{R}^n;$$

but this just means

$$\langle s, Ap \rangle \leq \sigma(p) \quad \text{for all } p \in \mathbb{R}^m,$$

i.e.  $A^*s \in S$ , because  $\langle s, Ap \rangle = \langle A^*s, p \rangle$ .  $\square$

Note that  $(A^*)^{-1}(S)$ , the inverse image of the closed set  $S$  under the continuous mapping  $A^*$ , is closed. By contrast,  $A\sigma$  need not be a closed function. As a particular case, suppose that  $S$  is bounded ( $\sigma_S$  is finite everywhere) and that  $A$  is surjective; then  $A\sigma$  is finite everywhere as well, which means that  $(A^*)^{-1}(S)$  is compact.

**Remark 3.3.6** The assumption made in Proposition 3.3.5 means exactly that the function  $A\sigma$  is nowhere  $-\infty$ ; in other words, its closure  $\text{cl}(A\sigma)$  is the support function of a *nonempty* set:  $(A^*)^{-1}(S) \neq \emptyset$ . This last property can be rewritten as

$$S \cap \text{Im } A^* \neq \emptyset \quad \text{or} \quad 0 \in S - \text{Im } A^* = S + (\text{Ker } A)^\perp. \quad (3.3.7)$$

On the other hand, the same starting assumption implies that (3.3.6) must hold in particular for  $d = 0$ . Then  $\sigma$  is bounded from below on the subspace  $\text{Ker } A$ ; by positive homogeneity, its lower bound is 0 on that subspace. Applying Corollary 3.3.2 and (3.3.3), this means

$$0 \in \text{cl}(P_{\text{Ker } A} S) \quad \text{or} \quad 0 \in \text{cl}[S + (\text{Ker } A)^\perp]. \quad (3.3.8)$$

Yet, the stronger property (3.3.7) is really necessary for Proposition 3.3.5 to hold. For a counter-example, define  $A : \mathbb{R}^2 \rightarrow \mathbb{R}$  by  $A(\xi, \eta) = \xi$  and  $S$  of (2.3.3). Then (3.3.8) holds but not (3.3.7); as seen in Example 2.3.2, we have here

$$(A\sigma_S)(\xi) = \inf_{\eta \leq 0} -2\sqrt{\xi\eta} \quad \text{for all } \xi \leq 0,$$

which is  $-\infty$  if  $\xi < 0$ .  $\square$

It has already been mentioned that taking an image-function is an important operation, from which several other operations can be constructed. We give two examples inspired from those at the end of §IV.2.4:

- Let  $S_1$  and  $S_2$  be two nonempty closed convex sets of  $\mathbb{R}^n$ , with support functions  $\sigma_1$  and  $\sigma_2$  respectively. With  $\mathbb{R}^m = \mathbb{R}^n \times \mathbb{R}^n$ , take  $A(x, y) = x + y$  and  $\sigma(d_1, d_2) = \sigma_1(d_1) + \sigma_2(d_2)$ ; observe that  $\sigma$  is the support function of  $S = S_1 \times S_2$ , associated with the scalar product

$$\langle\langle(s_1, s_2), (d_1, d_2)\rangle\rangle = \langle s_1, d_1 \rangle + \langle s_2, d_2 \rangle.$$

Then we obtain  $A\sigma = \sigma_1 \downarrow \sigma_2$ . On the other hand, the adjoint of  $A$  is clearly given by

$$A^*x = (x, x) \in \mathbb{R}^n \times \mathbb{R}^n \quad \text{for all } x \in \mathbb{R}^n,$$

so that the inverse image of  $S$  under  $A^*$  is nothing but  $S_1 \cap S_2$ : we recover (3.3.4).

- Let  $\sigma$  be the support function of some nonempty closed convex set  $S \subset \mathbb{R}^n \times \mathbb{R}^p$  and let  $A(x, y) = x$ , so that the image of  $\sigma$  under  $A$  is defined by

$$\mathbb{R}^n \ni x \mapsto (A\sigma)(x) = \inf \{\sigma(x, y) : y \in \mathbb{R}^p\}.$$

Now  $A^*$  is

$$\mathbb{R}^n \ni x \mapsto A^*x = (x, 0) \in \mathbb{R}^n \times \mathbb{R}^p$$

and we obtain that  $\text{cl}(A\sigma)$  is the support function of the “slice”

$$\{x \in \mathbb{R}^n : (x, 0) \in S\}.$$

This last set must not be confused with the projection of  $S$  onto  $\mathbb{R}^n$ , whose support function is  $x \mapsto \sigma_S(x, 0)$  (Proposition 3.3.4).

**Remark 3.3.7** Let us mention some more rules dealing with the operations reviewed in §IV.2.3:

- The closure of a perspective-function  $\tilde{f}$  is the support function of a (nonempty closed convex) set in  $\mathbb{R} \times \mathbb{R}^n$ . A good exercise is to try and figure out what it looks like; it will be extensively studied in Chap. XI (see §XI.1.2).
- The support function of a star-difference is obtained as follows. Let  $S_1$  and  $S_2$  be two nonempty closed convex sets, with  $S_2$  bounded; assuming  $S := S_1 \pm S_2 \neq \emptyset$ ,

$$\sigma_S = \overline{\text{co}}(\sigma_{S_1} - \sigma_{S_2}).$$

- On the other hand, the deconvolution yields an interesting exercise, even though the result is of little value. Let  $\sigma_1$  and  $\sigma_2$  be two closed sublinear functions. One can prove that their deconvolution  $\sigma_1 \bar{\vee} \sigma_2$  is closed, convex and positively homogeneous. Most of the time, it is identically  $+\infty$ ; a non-degenerate situation corresponds to  $(\sigma_1 \bar{\vee} \sigma_2)(0) = 0$ ; it is obtained if and only if  $\sigma_1 \leqslant \sigma_2$ , and then  $\sigma_1 \bar{\vee} \sigma_2 = \sigma_1$ .  $\square$

Having studied the isomorphism with respect to order and algebraic structures, we pass to topologies. Theorem 1.3.3 has defined a distance  $\Delta$  on the set of finite sublinear functions. Likewise, the *Hausdorff distance*  $\Delta_H$  can be defined for nonempty closed sets (see §A.5). When restricted to nonempty compact convex sets,  $\Delta_H$  plays the role of the distance introduced in Theorem 1.3.3:

**Theorem 3.3.8** *If  $S$  and  $S'$  are two nonempty compact convex sets of  $\mathbb{R}^n$ ,*

$$\Delta(\sigma_S, \sigma_{S'}) := \max_{\|d\| \leqslant 1} |\sigma_S(d) - \sigma_{S'}(d)| = \Delta_H(S, S'). \quad (3.3.9)$$

PROOF. As already mentioned, for all  $r \geq 0$ , the property

$$\max \{d_S(d) : d \in S'\} \leq r \quad (3.3.10)$$

simply means

$$S' \subset S + B(0, r).$$

Now, the support function of  $B(0, 1)$  is  $\|\cdot\|$  – see (2.3.1). Calculus rules on support functions therefore tell us that (3.3.10) is also equivalent to

$$\sigma_{S'}(d) \leq \sigma_S(d) + r\|d\| \quad \text{for all } d \in \mathbb{R}^n \iff \max_{\|d\| \leq 1} [\sigma_{S'}(d) - \sigma_S(d)] \leq r.$$

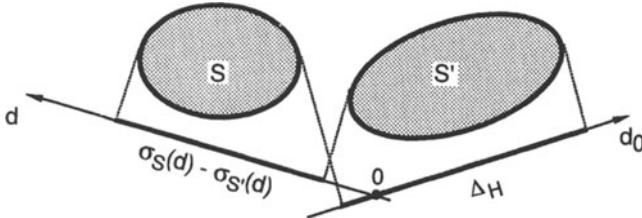
In summary, we have proved

$$\max_{d \in S'} d_S(d) = \max_{\|d\| \leq 1} [\sigma_{S'}(d) - \sigma_S(d)]$$

and symmetrically

$$\max_{d \in S} d_{S'}(d) = \max_{\|d\| \leq 1} [\sigma_S(d) - \sigma_{S'}(d)];$$

the result follows.  $\square$



**Fig. 3.3.1.** Hausdorff distances

Naturally, the max in (3.3.9) is attained at some  $d_0$ : for  $S$  and  $S'$  convex compact, there exists  $d_0$  of norm 1 such that

$$\Delta_H(S, S') = \Delta(\sigma_S, \sigma_{S'}) = |\sigma_S(d_0) - \sigma_{S'}(d_0)|.$$

Figure 3.3.1 illustrates a typical situation. When  $S' = \{0\}$ , we obtain the number

$$\Delta_H(\{0\}, S) = \max_{s \in S} \|s\| = \max_{\|d\|=1} \sigma_S(d),$$

already seen in (1.2.6); it is simply the distance from 0 to the most remote hyperplane  $H_{d, \sigma_S(d)}$  touching  $S$  (see again the end of Interpretation 2.1.5).

Using (3.3.9), it becomes rather easy to compute the distance in Example 1.3.4, which becomes the Hausdorff-distance (in fact an excess) between the corresponding unit balls.

When speaking of limits of nonempty convex compact sets to a nonempty convex compact set, the following result is a further illustration of our isomorphism.

**Proposition 3.3.9** *A convex-compact-valued and locally bounded multifunction  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is outer [resp. inner] semi-continuous at  $x_0 \in \text{int dom } F$  if and only if its support function  $x \mapsto \sigma_{F(x)}(d)$  is upper [resp. lower] semi-continuous at  $x_0$  for all  $d$  of norm 1.*

PROOF. Calculus with support functions tells us that our definition (A.5.2) of outer semi-continuity is equivalent to

$$\forall \varepsilon > 0, \exists \delta > 0 : y \in B(x_0, \delta) \implies \sigma_{F(y)}(d) \leq \sigma_{F(x_0)}(d) + \varepsilon \|d\| \text{ for all } d \in \mathbb{R}^n$$

and division by  $\|d\|$  shows that this is exactly upper semi-continuity of the support function for  $\|d\| = 1$ . Same proof for inner/lower semi-continuity.  $\square$

Thus, a convex-compact-valued, locally bounded mapping  $F$  is both outer and inner semi-continuous at  $x_0$  if and only if its support function  $\sigma_{F(\cdot)}(d)$  is continuous at  $x_0$  for all  $d$ . In view of Theorem 1.3.5,  $\sigma_{F(\cdot)}(d)$  is then continuous at  $x_0$  uniformly for  $d \in B(0, 1)$ ; and Theorem 3.3.8 tells us that this property in turn means:

$$\Delta_H(F(x), F(x_0)) \rightarrow 0 \quad \text{when } x \rightarrow x_0.$$

The following interpretation in terms of sequences is useful.

**Corollary 3.3.10** *Let  $\{S_k\}$  be a sequence of nonempty convex compact sets and  $S$  a nonempty convex compact set. When  $k \rightarrow +\infty$ , the following are equivalent*

- (i)  $S_k \rightarrow S$  in the Hausdorff sense, i.e.  $\Delta_H(S_k, S) \rightarrow 0$ ;
- (ii)  $\sigma_{S_k} \rightarrow \sigma_S$  pointwise;
- (iii)  $\sigma_{S_k} \rightarrow \sigma_S$  uniformly on each compact set of  $\mathbb{R}^n$ .

$\square$

Let us sum up this Section 3.3: when combining/comparing closed convex sets, one knows what happens to their support functions (apply the results 3.3.1 – 3.3.4). Conversely, when closed sublinear functions are combined/compared, one knows what happens to the sets they support. The various rules involved are summarized in Table 3.3.1. Each  $S_i$  is a nonempty closed convex set, with support function  $\sigma_i$ .

This table deserves some comments.

- Generally speaking, it helps to remember that when a set increases, its support function increases (first line); hence the “crossing” of closed convex hulls in the last two lines.
- The rule of the last line comes directly from the definition (2.1.1) if each  $S_i$  is thought of as a singleton.
- Most of these rules are still applicable without closed convexity of each  $S_i$  (remembering that  $\sigma_S = \sigma_{\text{co } S}$ ). For example, the equivalence in the first line requires closed convexity of  $S_2$  only. We mention one trap, however: when intersecting sets, each set must be closed and convex. A counter-example in one dimension is  $S_1 := \{0, 1\}$ ,  $S_2 := \{0, 2\}$ ; the support functions do not see the difference between  $S_1 \cap S_2$  and  $\text{co } S_1 \cap \text{co } S_2$ .

**Table 3.3.1.** Calculus rules for support functions

Closed convex sets	Closed sublinear functions
$S_1 \subset S_2$	$\sigma_1 \leq \sigma_2$
$\Delta_H(S_1, S_2)$ ( $S_i$ bounded)	$\Delta(\sigma_1, \sigma_2)$ ( $\sigma_i$ finite) uniform/compact or pointwise convergence (on finite functions)
Hausdorff convergence (on bounded sets)	
$tS$ ( $t > 0$ )	$t\sigma$
$\text{cl}(S_1 + S_2)$	$\sigma_1 + \sigma_2$
$\text{cl } A(S)$ ( $A$ linear)	$\sigma \circ A^*$
$(\bar{A}^*)(S)$ ( $A$ linear)	$\text{cl}(A\sigma)$
$\cap_{i \in I} S_i$ (nonempty)	$\overline{\text{co}} \inf_{i \in I} \sigma_i$ (minorized)
$\overline{\text{co}}(\cup_{i \in I} S_i)$	$\sup_{i \in I} \sigma_i$

**Example 3.3.11 (Maximal Eigenvalues)** Recall from §IV.1.3(e) that, if the eigenvalues of a symmetric matrix  $A$  are denoted by  $\lambda_1(A) \geq \dots \geq \lambda_n(A)$ , the function

$$\mathbf{S}_n(\mathbb{R}) \ni A \mapsto f_m(A) := \sum_{j=1}^m \lambda_j(A)$$

is convex – and finite everywhere. Its positive homogeneity is obvious, therefore it is the support function of a certain convex compact set  $C_m$  of symmetric matrices. Let us compute the set  $C_1$  when the scalar product in  $\mathbf{S}_n(\mathbb{R})$  is the standard dot-product of  $\mathbb{R}^{n \times n}$ :

$$\langle\langle A, B \rangle\rangle := \text{tr } AB = \sum_{i,j=1}^n A_{ij} B_{ij}.$$

Indeed, we know that

$$\lambda_1(A) = \sup_{x^\top x = 1} x^\top A x = \sup_{x^\top x = 1} \langle\langle x x^\top, A \rangle\rangle.$$

Hence  $C_1$  is the closed convex hull of the set of matrices

$$\{x x^\top : x^\top x = 1\},$$

which is clearly compact. Actually, its Hausdorff distance to  $\{0\}$  is

$$\Delta_H(\{0\}, C_1) = \max_{x^\top x = 1} \sqrt{\langle\langle x x^\top, x x^\top \rangle\rangle} = 1.$$

Incidentally,  $\lambda_1(\cdot)$  is therefore nonexpansive in  $\mathbf{S}_n(\mathbb{R})$ .

We leave it as an exercise to demonstrate the following nicer representation of  $C_1$ :

$$C_1 = \text{co } \{x x^\top : x^\top x = 1\} = \{M \in \mathbf{S}_n(\mathbb{R}) : \lambda_n(M) \geq 0, \text{tr } M = 1\}. \quad \square$$

### 3.4 Example: Support Functions of Closed Convex Polyhedra

In optimization, polyhedral sets are encountered all the time, and thus deserve special study. They are often defined by finitely many affine constraints, i.e. obtained as intersections of closed half-spaces; in view of Table 3.3.1, this explains that the infimal convolution encountered in Proposition 1.3.2 is fairly important.

**Example 3.4.1 (Compact Convex Polyhedra)** First of all, the support function of a polyhedron defined as

$$P := \text{co}\{p_1, \dots, p_m\} \quad (3.4.1)$$

is trivially

$$d \mapsto \sigma_P(d) = \max \{\langle p_i, d \rangle : i = 1, \dots, m\}.$$

There is no need to invoke Theorem 3.3.3 for this: a linear function  $\langle \cdot, d \rangle$  attains its maximum on an extreme point of  $P$  (Proposition III.2.4.6), even if this extreme point is not the entire face exposed by  $d$ .  $\square$

**Example 3.4.2 (Closed Convex Polyhedral Cones)** Going back to Example 2.3.1, suppose that the cone  $K$  is given as a finite intersection of half-spaces:

$$K = \cap \{K_j : j = 1, \dots, m\}, \quad (3.4.2)$$

where

$$K_j := H_{a_j, 0}^- := \{s \in \mathbb{R}^n : \langle a_j, s \rangle \leq 0\} \quad (3.4.3)$$

(the  $a_j$ 's are assumed nonzero). We use Proposition 1.3.2:

$$\sigma_K(d) = \text{cl inf} \left\{ \sum_{j=1}^m \sigma_{K_j}(d_j) : \sum_{j=1}^m d_j = d \right\}.$$

Only those  $d_j$  in  $K_j^\circ$  – namely nonnegative multiples of  $a_j$ , see (2.3.2) – count to yield the infimum; their corresponding support vanishes and we obtain

$$\sigma_K(d) = \begin{cases} 0 & \text{if } d = \sum_{j=1}^m t_j a_j, \ t_j \geq 0 \text{ for } j = 1, \dots, m, \\ +\infty & \text{otherwise.} \end{cases}$$

Here, we are lucky: the closure operation is useless because the right-hand side is already a closed convex function. Note that we recognize Farkas' Lemma III.4.3.3:  $K^\circ = \text{dom } \sigma_K$  is the conical hull of the  $a_j$ 's, which is closed thanks to the fact that there are *finitely many* generators.  $\square$

**Example 3.4.3 (Extreme Points and Directions)** Suppose our polyhedron is defined in the spirit of 3.4.1, but unbounded:

$$S := \text{co}\{p_1, \dots, p_m\} + \text{cone}\{a_1, \dots, a_\ell\}.$$

Then it suffices to observe that  $S = P + K^\circ$ , with  $P$  of (3.4.1) and  $K$  of (3.4.2), (3.4.3). Using Table 3.3.1 and knowing that  $K^{\circ\circ} = K$  – hence  $\sigma_{K^\circ} = I_K$ :

$$\sigma_S(d) = \begin{cases} \max_{i=1, \dots, m} \langle p_i, d \rangle & \text{if } \langle a_j, d \rangle \leq 0 \text{ for } j = 1, \dots, \ell, \\ +\infty & \text{otherwise.} \end{cases}$$

$\square$

The representations of Examples 3.4.1 and 3.4.3 are not encountered so frequently. Our next examples, dealing with intersections, represent the vast majority of situations.

**Example 3.4.4 (Inequality Constraints)** Perturb Example 3.4.2 to express the support function of  $S := \cap H_{a_j, b_j}^-$ , with

$$H_{a, b}^- := \{s \in \mathbb{R}^n : \langle s, a \rangle \leq b\} \quad (a \neq 0).$$

Here, we deal with translations of the  $K_j$ 's:

$$H_{a_j, b_j}^- = \frac{b_j}{\|a_j\|^2} a_j + K_j$$

so, with the help of Table 3.3.1:

$$\sigma_{H_{a_j, b_j}^-}(d) = \begin{cases} tb_j & \text{if } d = ta_j, t \geq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Provided that  $S \neq \emptyset$ , our support function  $\sigma_S$  is therefore the closure of the function

$$d \mapsto \begin{cases} \inf \left\{ \sum_{j=1}^m t_j b_j : \sum_{j=1}^m t_j a_j = d, t_j \geq 0 \right\} & \text{if } d \in \text{cone}(a_1, \dots, a_m), \\ +\infty & \text{otherwise.} \end{cases} \quad \square$$

Now we have a sudden complication: the domain of  $\sigma$  is still the closed convex cone  $K^\circ$ , but the status of the closure operation is no longer quite clear. Also, it is not even clear whether the above infimum is attained. Actually, all this results from Farkas' Lemma of §III.4.3; before giving the details, let us adopt different notations.

**Example 3.4.5 (Closed Convex Polyhedra in Standard Form)** Equalities can be formulated as pairs of reversed inequalities, thus enabling Example 3.4.4 to treat any kind of constraints. A “standard” description of closed convex polyhedra, however, is as follows. Let  $A$  be a linear operator from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ ,  $b \in \text{Im } A \subset \mathbb{R}^m$ ,  $K \subset \mathbb{R}^n$  a closed convex polyhedral cone ( $K$  is usually characterized as in Example 3.4.2). Then  $S$  is given by

$$S := \{s \in \mathbb{R}^n : As = b, s \in K\} = (\{s_0\} + H) \cap K, \quad (3.4.4)$$

where  $s_0$  is some point in  $\mathbb{R}^n$  satisfying  $As_0 = b$ , and  $H := \text{Ker } A$ .

In view of the expression of  $\sigma_H$  in Example 2.3.1, the support function of  $\{s_0\} + H$  is finite only on  $\text{Im } A^*$ , where it is equal to

$$\sigma_{\{s_0\}}(d) + \sigma_H(d) = \langle s_0, d \rangle = \langle b, z \rangle \quad \text{for } d = A^*z, z \in \mathbb{R}^m$$

(here,  $\langle \cdot, \cdot \rangle$  denotes the scalar product in  $\mathbb{R}^m$ ). Thus,  $\sigma_S$  is the closure of the infimal convolution

$$(\sigma_{\{s_0\}} + \sigma_H) \downarrow \sigma_K = (\sigma_{\{s_0\}} + \sigma_H) \downarrow I_{K^\circ}, \quad (3.4.5)$$

which can be explicated as the function

$$d \mapsto \inf \{ \langle \langle b, z \rangle \rangle : (z, y) \in \mathbb{R}^m \times K^\circ, A^*z + y = d \}.$$

Of course, this formula clearly displays

$$\text{dom } \sigma_S = \text{dom } \sigma_H + \text{dom } I_{K^\circ} = \text{Im } A^* + K^\circ.$$

In the pure *standard form*,  $\mathbb{R}^n$  and  $\mathbb{R}^m$  are both equipped with the standard dot-product –  $A$  being a matrix with  $m$  rows and  $n$  columns – and  $K$  is the nonnegative orthant;  $K^\circ$  is therefore the nonpositive orthant. Our “standard”  $S$  of (3.4.4) is now

$$\{s \in \mathbb{R}^n : As = b, s \geq 0\}, \quad (3.4.6)$$

assumed nonempty. Then (3.4.5) becomes

$$\inf \{b^\top z : A^\top z \geq d\}, \quad (3.4.7)$$

a function of  $d$  which is by no means simpler than in 3.4.4 – only the notation is different. In summary, the support function

$$\sigma_S(d) = \sup \{s^\top d : As = b, s \geq 0\} \quad (3.4.8)$$

of the set (3.4.6) is the closure of (3.4.7), considered as a function of  $d \in \mathbb{R}^n$ .

Now, invoke Farkas’ Lemma: write the equivalent statements (i)” and (ii)” from the end of §III.4, with  $(x, \rho, \alpha, r)$  changed to  $(-z, -d, s, -\sigma)$ :

$$\{z \in \mathbb{R}^n : A^\top z \geq d\} \subset \{z \in \mathbb{R}^n : b^\top z \geq \sigma\} \quad (3.4.9)$$

is equivalent to

$$\exists s \geq 0 \text{ such that } As = b, s^\top d \geq \sigma. \quad (3.4.10)$$

In other words: the largest  $\sigma$  for which (3.4.9) holds – i.e. the value (3.4.7) – is also the largest  $\sigma$  for which (3.4.10) holds – i.e.  $\sigma_S(d)$ . The closure operation can be omitted and we do have

$$\sigma_S(d) = \inf \{b^\top z : A^\top z \geq d\} \text{ for all } d \in \mathbb{R}^n.$$

Another interesting consequence can also be noted. Take  $d$  such that  $\sigma_S(d) < +\infty$ : if we put  $\sigma = \sigma_S(d)$  in (3.4.9), we obtain a true statement, i.e. (3.4.10) is also true. This means that the supremum in (3.4.8) is attained when it is finite.  $\square$

It is worth noting that Example 3.4.5 describes general polyhedral functions, up to notational changes. As such, it discloses results of general interest, namely:

- A linear function which is bounded from above on a closed convex polyhedron attains its maximum on this polyhedron.
- The infimum of a linear function under affine constraints is a closed sublinear function of the right-hand side; said otherwise, an image of a polyhedral function is closed: in Example 3.4.5, the polyhedral function in question is

$$\mathbb{R}^m \times \mathbb{R}^n \ni (y, z) \mapsto b^\top z + I_K(y),$$

and (3.4.7) gives its image under the linear mapping  $[A^\top | 0]$ .

## VI. Subdifferentials of Finite Convex Functions

**Prerequisites.** First-order differentiation of convex functions of one real variable (Chap. I); basic definitions, properties and operations concerning finite convex functions (Chap. IV); finite sublinear functions and support functions of compact convex sets (Chap. V).

**Introduction.** We have mentioned in our preamble to Chap. V that sublinearity permits the approximation of convex functions to first order around a given point. In fact, we will show here that, if  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is convex and  $x \in \mathbb{R}^n$  is fixed, then the function

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

exists and is finite *sublinear*. Furthermore,  $f'$  approximates  $f$  around  $x$  in the sense that

$$f(x + h) = f(x) + f'(x, h) + o(\|h\|). \quad (0.1)$$

In view of the correspondence between finite sublinear functions and compact convex sets (which formed a large part of Chap. V),  $f'(x, \cdot)$  can be expressed for all  $d \in \mathbb{R}^n$  as

$$f'(x, d) = \sigma_S(d) = \max \{\langle s, d \rangle : s \in S\}$$

for some nonempty compact convex set  $S$ . This  $S$  is called the *subdifferential* of  $f$  at  $x$  and is traditionally denoted by  $\partial f(x)$ . When  $f$  is differentiable at  $x$ , with gradient  $\nabla f(x)$ , (0.1) shows that  $f'(x, \cdot)$  becomes linear and  $S$  contains only the element  $\nabla f(x)$ . Thus, the concept of subdifferential generalizes that of gradient, just as sublinearity generalizes linearity.

This subdifferential has already been encountered in the one-dimensional case where  $x \in \mathbb{R}$ . In that context,  $\partial f(x)$  was a closed interval, interpreted as a set of slopes. Its properties developed there will be reconsidered here, using the powerful apparatus of support functions studied in Chap. V.

The subdifferentiation thus introduced is supposed to generalize the ordinary differentiation; one should therefore not be surprised to find counterparts of most of the results encountered in differential calculus: first-order Taylor expansions, mean-value theorems, calculus rules, etc. The importance of calculus rules increases in the framework of convex analysis: some operations on convex functions destroy differentiability (and thereby find no place in differential calculus) but preserve convexity. An important example is the max-operation; indeed, we will give a detailed account of the calculus rules for the subdifferential of max-functions.

This chapter deals with *finite-valued* convex functions exclusively: it is essential for practitioners to have a good command of subdifferential calculus, and this framework is good enough. Furthermore, its generalization to the extended-valued case (Chap. XI) will be easier to assimilate. Unless otherwise specified, therefore:

$$f : \mathbb{R}^n \rightarrow \mathbb{R} \text{ is convex .}$$

This implies the continuity and local Lipschitz continuity of  $f$ . We note from (0.1), however, that the concept of subdifferential is essentially local; for an extended-valued  $f$ , most results in this chapter remain true at a point  $x \in \text{int dom } f$  (assumed non-empty). It can be considered as an exercise to check those results in this generalized setting – with answers given in Chap. XI.

## 1 The Subdifferential: Definitions and Interpretations

### 1.1 First Definition: Directional Derivatives

Let  $x$  and  $d$  be fixed in  $\mathbb{R}^n$  and consider the difference quotient of  $f$  at  $x$  in the direction  $d$ :

$$q(t) := \frac{f(x + td) - f(x)}{t} \quad \text{for } t > 0. \quad (1.1.1)$$

We have seen already that the function  $t \mapsto q(t)$  is increasing (criterion I.1.1.4 of increasing slopes) and bounded near 0 (local Lipschitz property of  $f$ , §IV.3.1); so the following definition makes sense.

**Definition 1.1.1** The *directional derivative* of  $f$  at  $x$  in the direction  $d$  is

$$f'(x, d) := \lim_{t \downarrow 0} \{q(t) : t \downarrow 0\} = \inf \{q(t) : t > 0\}. \quad (1.1.2)$$

□

If  $\varphi$  denotes the one-dimensional function  $t \mapsto \varphi(t) := f(x + td)$ , then

$$f'(x, d) = D_+ \varphi(0) \quad (1.1.3)$$

is nothing other than the right-derivative of  $\varphi$  at 0 (see §I.4.1). Changing  $d$  to  $-d$  in (1.1.1), one obtains

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

which is *not* the left-derivative of  $\varphi$  at 0 but rather its negative counterpart:

$$f'(x, -d) = -D_- \varphi(0). \quad (1.1.4)$$

**Proposition 1.1.2** For fixed  $x$ , the function  $f'(x, \cdot)$  is finite sublinear.

PROOF. Let  $d_1, d_2$  in  $\mathbb{R}^n$ , and positive  $\alpha_1, \alpha_2$  with  $\alpha_1 + \alpha_2 = 1$ . From the convexity of  $f$ :

$$\begin{aligned} f(x + t(\alpha_1 d_1 + \alpha_2 d_2)) - f(x) &= \\ f(\alpha_1(x + td_1) + \alpha_2(x + td_2)) - \alpha_1 f(x) - \alpha_2 f(x) &\leqslant \\ \leqslant \alpha_1[f(x + td_1) - f(x)] + \alpha_2[f(x + td_2) - f(x)] \end{aligned}$$

for all  $t$ . Dividing by  $t > 0$  and letting  $t \downarrow 0$ , we obtain

$$f'(x, \alpha_1 d_1 + \alpha_2 d_2) \leqslant \alpha_1 f'(x, d_1) + \alpha_2 f'(x, d_2)$$

which establishes the convexity of  $f'$  with respect to  $d$ . Its positive homogeneity is clear: for  $\lambda > 0$

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

Finally suppose  $\|d\| = 1$ . As a finite convex function,  $f$  is Lipschitz continuous around  $x$  (Theorem IV.3.1.2); in particular there exist  $\varepsilon > 0$  and  $L > 0$  such that

$$|f(x + t d) - f(x)| \leqslant L t \quad \text{for } 0 \leqslant t \leqslant \varepsilon.$$

Hence,  $|f'(x, d)| \leqslant L$  and we conclude with positive homogeneity:

$$|f'(x, d)| \leqslant L \|d\| \quad \text{for all } d \in \mathbb{R}^n. \quad (1.1.5)$$

□

**Remark 1.1.3** From the end of the above proof, a local Lipschitz constant  $L$  of  $f$  around  $x$  is transferred to  $f'(\cdot, \cdot)$  via (1.1.5). In view of (V.1.2.6), this same  $L$  is a *global* Lipschitz constant for  $f'(\cdot, \cdot)$ . This is even true of  $f'(y, \cdot)$  for  $y$  close to  $x$ : with  $\delta$  and  $L$  such that  $f$  has the Lipschitz constant  $L$  on  $B(x, \delta)$ ,

$$\|y - x\| < \delta \implies |f'(y, d_1) - f'(y, d_2)| \leqslant L \|d_1 - d_2\| \quad \text{for all } d_1, d_2 \in \mathbb{R}^n. \quad \square$$

A consequence of Proposition 1.1.2 is that  $f'(\cdot, \cdot)$  is a support function, so the following suggests itself:

**Definition 1.1.4 (Subdifferential I)** The subdifferential  $\partial f(x)$  of  $f$  at  $x$  is the non-empty compact convex set of  $\mathbb{R}^n$  whose support function is  $f'(\cdot, \cdot)$ , i.e.

$$\partial f(x) := \{s \in \mathbb{R}^n : \langle s, d \rangle \leqslant f'(x, d) \text{ for all } d \in \mathbb{R}^n\}. \quad (1.1.6)$$

A vector  $s \in \partial f(x)$  is called a *subgradient* of  $f$  at  $x$ .

□

A first observation is therefore that the concept of subdifferential is *attached* to a scalar product, just because the concept of support is so. All the properties of the correspondence between compact convex sets and finite sublinear functions can be reformulated for  $\partial f(x)$  and  $f'(\cdot, \cdot)$ . For example, the breadth of  $\partial f(x)$  (cf. Definition V.2.1.4) along a normalized direction  $d$  is

$$f'(x, d) + f'(x, -d) = D_+ \varphi(0) - D_- \varphi(0) \geqslant 0$$

and represents the “lack of differentiability” of the function  $\varphi$  alluded to in (1.1.3), (1.1.4); remember Proposition IV.4.2.1.

**Remark 1.1.5** In particular, for  $d$  in the subspace  $U$  of linearity of  $f'(x, \cdot)$  – see (V.1.1.8) – the corresponding  $\varphi$  is differentiable at 0. The restriction of  $f'(x, \cdot)$  to  $U$  is linear (Proposition V.1.1.6) and equals  $\langle s, \cdot \rangle$ , no matter how  $s$  is chosen in  $\partial f(x)$ . In words,  $U$  is the set of  $h$  for which  $h \mapsto f(x + h)$  behaves as a function differentiable at  $h = 0$ . See Fig. 1.1.1:  $\partial f(x)$  is entirely contained in a hyperplane parallel to  $U$ ; said otherwise,  $U$  is the set of directions along which  $\partial f(x)$  has 0-breadth.

We also recall from Definition V.2.1.4 that

$$-f'(x, -d) \leq \langle s, d \rangle \leq f'(x, d) \quad \text{for all } (s, d) \in \partial f(x) \times \mathbb{R}^n.$$

□

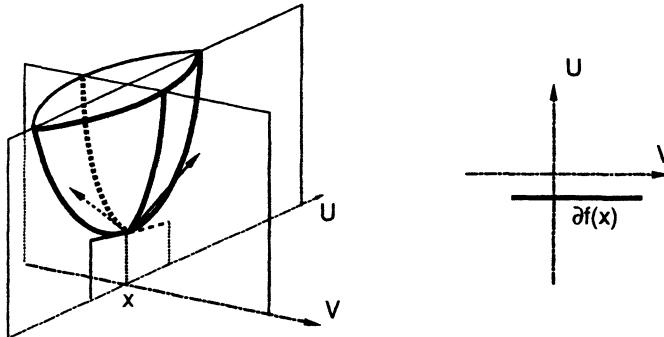


Fig. 1.1.1. Linearity-space of the directional derivative

It results directly from Chap. V that Definition 1.1.4 can also be looked at from the other side: (1.1.6) is equivalent to

$$f'(x, d) = \sup \{ \langle s, d \rangle : s \in \partial f(x) \}.$$

Remembering that  $\partial f(x)$  is compact, this supremum is attained at some  $s$  – which depends on  $d$ ! In other words: for any  $d \in \mathbb{R}^n$ , there is some  $s_d \in \partial f(x)$  such that

$$f(x + td) = f(x) + t\langle s_d, d \rangle + t\varepsilon_d(t) \quad \text{for } t \geq 0. \quad (1.1.7)$$

Here  $\varepsilon_d(t) \rightarrow 0$  for  $t \downarrow 0$ , and we will see later that  $\varepsilon_d$  can actually be made independent of the normalized  $d$ ; as for  $s_d$ , it is a subgradient giving the largest  $\langle s, d \rangle$ . Thus, from its very construction, the subdifferential contains all the necessary information for a first-order description of  $f$ .

As a finite convex function,  $d \mapsto f'(x, d)$  has itself directional derivatives and subdifferentials. These objects at  $d = 0$  are of particular interest; the case  $d \neq 0$  will be considered later.

**Proposition 1.1.6** *The finite sublinear function  $d \mapsto \sigma(d) := f'(x, d)$  satisfies*

$$\sigma'(0, \delta) = f'(x, \delta) \quad \text{for all } \delta \in \mathbb{R}^n; \quad (1.1.8)$$

$$\sigma(\delta) = \sigma(0) + \sigma'(0, \delta) = \sigma'(0, \delta) \quad \text{for all } \delta \in \mathbb{R}^n; \quad (1.1.9)$$

$$\partial\sigma(0) = \partial f(x). \quad (1.1.10)$$

PROOF. Because  $\sigma$  is positively homogeneous and  $\sigma(0) = 0$ ,

$$\frac{\sigma(t\delta) - \sigma(0)}{t} = \sigma(\delta) = f'(x, \delta) \quad \text{for all } t > 0.$$

This implies immediately (1.1.8) and (1.1.9). Then (1.1.10) follows from uniqueness of the supported set.  $\square$

One should not be astonished by (1.1.8): tangency is a self-reproducing operation. Since the graph of  $f'(x, \cdot)$  is made up of the (half-)lines tangent to  $\text{gr } f$  at  $(x, f(x))$ , the same set must be obtained when taking the (half-)lines tangent to  $\text{gr } f'(x, \cdot)$ . As for (1.1.9), it simply expresses that, when developing a sublinear function to first order at 0, there is no error of linearization: (1.1.7) holds with  $\varepsilon_d \equiv 0$  in that case.

## 1.2 Second Definition: Minorization by Affine Functions

The previous Definition 1.1.4 of the subdifferential involved two steps: first, calculating the directional derivative, and then determining the set that it supports. It is however possible to give a direct definition, with no reference to differentiation.

**Definition 1.2.1 (Subdifferential II)** The subdifferential of  $f$  at  $x$  is the set of vectors  $s \in \mathbb{R}^n$  satisfying

$$f(y) \geq f(x) + \langle s, y - x \rangle \quad \text{for all } y \in \mathbb{R}^n. \quad (1.2.1)$$

$\square$

Of course, we have to prove that our new definition coincides with 1.1.4. This will be done in Theorem 1.2.2 below. First, we make a few remarks illustrating the difference between Definitions 1.1.4 and 1.2.1.

- The present definition is *unilateral*: an inequality is required in (1.2.1), expressing the fact that the affine function  $y \mapsto f(x) + \langle s, y - x \rangle$  *minorizes*  $f$  and *coincides* with  $f$  for  $y = x$ .
- It is a *global* definition, in the sense that (1.2.1) involves all  $y$  in  $\mathbb{R}^n$ .
- These two observations do suggest that 1.2.1 deviates from the concept of differentiation, namely:

- (i) no remainder term shows up in (1.2.1), and
- (ii) every  $y$  counts, not only those close to  $x$ .

Actually, the proof below will show that nothing changes if:

- (i') an extra  $o(\|y - x\|)$  is added in (1.2.1), or
- (ii') (1.2.1) is required to hold for  $y$  close to  $x$  only.

Of course, these two properties (i') and (ii') rely on convexity of  $f$ ; more precisely on monotonicity of the difference quotient.

- All subgradients are described by (1.2.1) at the same time. By contrast,  $f'(x, d) = \langle s_d, d \rangle$  plots, for  $d \neq 0$ , only the boundary of  $\partial f(x)$ , one exposed face at a time. The whole subdifferential is then obtained by convexification – remember Proposition V.3.1.5.

**Theorem 1.2.2** *The definitions 1.1.4 and 1.2.1 are equivalent.*

PROOF. Let  $s$  satisfy (1.1.6), i.e.

$$\langle s, d \rangle \leq f'(x, d) \quad \text{for all } d \in \mathbb{R}^n. \quad (1.2.2)$$

The second equality in (1.1.2) makes it clear that (1.2.2) is equivalent to

$$\langle s, d \rangle \leq \frac{f(x + td) - f(x)}{t} \quad \text{for all } d \in \mathbb{R}^n \text{ and } t > 0. \quad (1.2.3)$$

When  $d$  describes  $\mathbb{R}^n$  and  $t$  describes  $\mathbb{R}_*$ ,  $y := x + td$  describes  $\mathbb{R}^n$  and we realize that (1.2.3) is just (1.2.1).  $\square$

The above proof is deeper than it looks: because of the monotonicity of slopes, the inequality of (1.2.3) holds whenever it holds for all  $(d, t) \in B(0, 1) \times ]0, \varepsilon]$ . Alternatively, this means that nothing is changed in (1.2.1) if  $y$  is restricted to a neighborhood of  $x$ .

It is interesting to note that, in terms of first-order approximation of  $f$ , (1.2.1) brings some additional information to (1.1.7): it says that the remainder term  $\varepsilon_d(t)$  is nonnegative for all  $t \geq 0$ . On the other hand, (1.1.7) says that, for some specific  $s$  (depending on  $y$ ), (1.2.1) holds almost as an equality for  $y$  close to  $x$ .

Now, the path “directional derivative  $\rightarrow$  subdifferential” adopted in §1.1 can be reproduced backwards: the set defined in (1.2.1) is

- nonempty (Proposition IV.1.2.1),
- closed and convex (immediate from the definitions),
- bounded, due to a simple Lipschitz argument: for given  $0 \neq s \in \partial f(x)$ , take in (1.2.1)  $y = x + \delta s / \|s\|$  ( $\delta > 0$  arbitrary) to obtain

$$f(x) + L\delta \geq f(y) \geq f(x) + \delta \|s\|,$$

where the first inequality comes from the Lipschitz property IV.3.1.2, written on the compact set  $B(x, \delta)$ .

As a result, this set of (1.2.1) has a finite-valued support function. Theorem 1.2.2 simply tells us that this support function is precisely the directional derivative  $f'(x, \cdot)$  of (1.1.2).

**Remark 1.2.3** A (finite) sublinear function  $\sigma$  has a subdifferential, just as any other convex function. Its subdifferential at 0 is defined by

$$\partial\sigma(0) = \{s : \langle d, s \rangle \leq \sigma(d) \text{ for all } d \in \mathbb{R}^n\},$$

in which we recognize Theorem V.2.2.2. This permits a more compact way than (V.2.2.2) to construct a set from its support function: a (finite) sublinear function is the support of its subdifferential at 0.

In Fig. V.2.2.1, for example, the wording “filter with  $\langle s, \cdot \rangle \leq \sigma$ ?” can be replaced by the more elegant “take the subdifferential of  $\sigma$  at 0”.  $\square$

### 1.3 Geometric Constructions and Interpretations

Definition 1.2.1 means that the elements of  $\partial f(x)$  are the slopes of the hyperplanes supporting the epigraph of  $f$  at  $(x, f(x)) \in \mathbb{R}^n \times \mathbb{R}$ . In terms of tangent and normal cones, this is expressed by the following result, which could serve as a third definition of the subdifferential and directional derivative.

#### Proposition 1.3.1

- (i) A vector  $s \in \mathbb{R}^n$  is a subgradient of  $f$  at  $x$  if and only if  $(s, -1) \in \mathbb{R}^n \times \mathbb{R}$  is normal to  $\text{epi } f$  at  $(x, f(x))$ . In other words:

$$N_{\text{epi } f}(x, f(x)) = \{(\lambda s, -\lambda) : s \in \partial f(x), \lambda \geq 0\}.$$

- (ii) The tangent cone to the set  $\text{epi } f$  at  $(x, f(x))$  is the epigraph of the directional-derivative function  $d \mapsto f'(x, d)$ :

$$T_{\text{epi } f}(x, f(x)) = \{(d, r) : r \geq f'(x, d)\}.$$

PROOF. [(i)] Apply Definition III.5.2.3 to see that  $(s, -1) \in N_{\text{epi } f}(x, f(x))$  means

$$\langle s, y - x \rangle + (-1)[r - f(x)] \leq 0 \quad \text{for all } y \in \mathbb{R}^n \text{ and } r \geq f(y)$$

and the equivalence with (1.2.1) is clear. The formula follows since the set of normals forms a cone containing the origin.

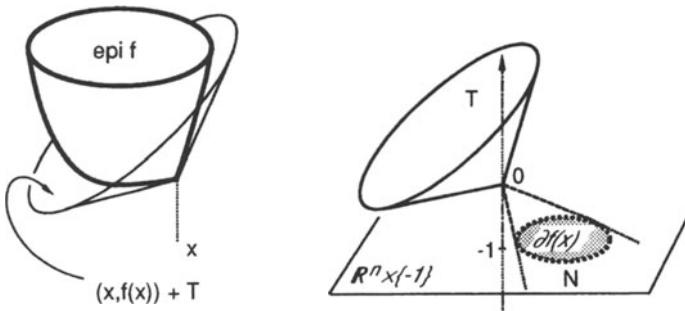
- [(ii)] The tangent cone to  $\text{epi } f$  is the polar of the above normal cone, i.e. the set of  $(d, r) \in \mathbb{R}^n \times \mathbb{R}$  such that

$$\langle \lambda s, d \rangle + (-\lambda)r \leq 0 \quad \text{for all } s \in \partial f(x) \text{ and } \lambda \geq 0.$$

Barring the trivial case  $\lambda = 0$ , we divide by  $\lambda > 0$  to obtain

$$r \geq \max \{\langle s, d \rangle : s \in \partial f(x)\} = f'(x, d).$$

□



**Fig. 1.3.1.** Tangents and normals to the epigraph

Figure 1.3.1 illustrates this result. The right part of the picture represents the normal cone  $N$  and tangent cone  $T$  to  $\text{epi } f$  at  $(x, f(x))$ . The intersection of  $N$  with

the space  $\mathbb{R}^n$  at level  $-1$  is just  $\partial f(x) \times \{-1\}$ . On the left part of the picture, the origin is translated to  $(x, f(x))$  and the translated  $T$  is tangent to  $\text{epi } f$ . Note that the boundary of  $T + (x, f(x))$  is also a (nonconvex) cone, “tangent”, in the intuitive sense of the term, to the graph of  $f$ ; it is the graph of  $f'(x, \cdot)$ , translated at  $(x, f(x))$ .

Proposition 1.3.1 and its associated Fig. 1.3.1 refer to Interpretation V2.1.6, with a supported set drawn in  $\mathbb{R}^n \times \mathbb{R}$ . One can also use Interpretation V2.1.5, in which the supported set was drawn in  $\mathbb{R}^n$ . In this framework, the sublevel-set passing through  $x$

$$Sf(x) := S_{f(x)}(f) = \{y \in \mathbb{R}^n : f(y) \leq f(x)\} \quad (1.3.1)$$

is particularly interesting: it is important for minimization, and it is closely related to  $\partial f(x)$ .

**Lemma 1.3.2** *For the convex function  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  and the sublevel-set (1.3.1), we have*

$$T_{Sf(x)}(x) \subset \{d : f'(x, d) \leq 0\}. \quad (1.3.2)$$

PROOF. Take arbitrary  $y \in Sf(x)$ ,  $t > 0$ , and set  $d := t(y - x)$ . Then, using the second equality in (1.1.2),

$$0 \geq t[f(y) - f(x)] = \frac{f(x + d/t) - f(x)}{1/t} \geq f'(x, d).$$

So we have proved

$$\mathbb{R}^+ [Sf(x) - x] \subset \{d : f'(x, d) \leq 0\} \quad (1.3.3)$$

(note: the case  $d = 0$  is covered since  $0 \in Sf(x) - x$ ).

Because  $f'(x, \cdot)$  is a closed function, the right-hand set in (1.3.3) is closed. Knowing that  $T_{Sf(x)}(x)$  is the closure of the left-hand side in (1.3.3) (Proposition III.5.2.1), we deduce the result by taking the closure of both sides in (1.3.3).  $\square$

The reader should be warned that the converse inclusion in (1.3.2) need not hold: for a counter-example, take  $f(x) = 1/2 \|x\|^2$ . The sublevel-set  $Sf(0)$  is then  $\{0\}$  and  $f'(0, d) = 0$  for all  $d$ . In this case, (1.3.2) reads  $\{0\} \subset \mathbb{R}^n$ ! To prove the converse inclusion, an additional assumption is definitely needed – for example the one considered in the following technical result.

**Proposition 1.3.3** *Let  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  be convex and suppose that  $g(x_0) < 0$  for some  $x_0 \in \mathbb{R}^n$ . Then*

$$\text{cl} \{z : g(z) < 0\} = \{z : g(z) \leq 0\}, \quad (1.3.4)$$

$$\{z : g(z) < 0\} = \text{int} \{z : g(z) \leq 0\}. \quad (1.3.5)$$

*It follows*

$$\text{bd} \{z : g(z) \leq 0\} = \{z : g(z) = 0\}. \quad (1.3.6)$$

PROOF. Because  $g$  is (lower semi-) continuous, the inclusion “ $\subset$ ” automatically holds in (1.3.4). Conversely, let  $\bar{z}$  be arbitrary with  $g(\bar{z}) \leq 0$  and, for  $k > 0$ , set

$$z_k := \frac{1}{k}x_0 + (1 - \frac{1}{k})\bar{z}.$$

By convexity of  $g$ ,  $g(z_k) < 0$ , so (1.3.4) is established by letting  $k \rightarrow +\infty$ .

Now, take the interior of both sides in (1.3.4). The “int cl” on the left is actually an “int” (Proposition III.2.1.8), and this “int”-operation is useless because  $g$  is (upper semi-) continuous: (1.3.5) is established.  $\square$

The existence of  $x_0$  in this result is often called a *Slater assumption*, and will be useful in the next chapters. When this  $x_0$  exists, taking closures, interiors and boundaries of sublevel-sets amounts to imposing “ $\leq$ ”, “ $<$ ” and “ $=$ ” in their definitions. Needless to say, convexity is essential for such an equivalence: with  $n = 1$ , think of  $g(z) := \min\{0, |z| - 1\}$ .

We are now in a position to characterize the tangential elements to a sublevel-set.

**Theorem 1.3.4** *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be convex and suppose  $0 \notin \partial f(x)$ . Then,  $Sf(x)$  being the sublevel-set (1.3.1),*

$$T_{Sf(x)}(x) = \{d \in \mathbb{R}^n : f'(x, d) \leq 0\} \quad (1.3.7)$$

$$\text{int}[T_{Sf(x)}(x)] = \{d \in \mathbb{R}^n : f'(x, d) < 0\} \neq \emptyset. \quad (1.3.8)$$

PROOF. From the very definition (1.1.6), our assumption means that  $f'(x, d) < 0$  for some  $d$ , and (1.1.2) then implies that  $f(x + td) < f(x)$  for  $t > 0$  small enough: our  $d$  is of the form  $(x + td - x)/t$  with  $x + td \in Sf(x)$  and we have proved

$$\{d : f'(x, d) < 0\} \subset \mathbb{R}^+ [Sf(x) - x] \subset T_{Sf(x)}(x). \quad (1.3.9)$$

Now, we can apply (1.3.4) with  $g = f'(x, \cdot)$ :

$$\text{cl}\{d : f'(x, d) < 0\} = \{d : f'(x, d) \leq 0\},$$

so (1.3.7) is proved by closing the sets in (1.3.9) and using (1.3.2). Finally, take the interior of both sides in (1.3.7) and apply (1.3.5) with  $g = f'(x, \cdot)$  to prove (1.3.8).  $\square$

The above result can be formulated in terms of normal cones.

**Theorem 1.3.5** *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be convex and suppose  $0 \notin \partial f(x)$ . Then a direction  $d$  is normal to  $Sf(x)$  at  $x$  if and only if there is some  $t \geq 0$  and some  $s \in \partial f(x)$  such that  $d = ts$ :*

$$N_{Sf(x)}(x) = \mathbb{R}^+ \partial f(x).$$

PROOF. Write (1.3.7) as

$$\begin{aligned} T_{Sf(x)}(x) &= \{d \in \mathbb{R}^n : \langle s, d \rangle \leq 0 \text{ for all } s \in \partial f(x)\} \\ &= \{d \in \mathbb{R}^n : \langle \lambda s, d \rangle \leq 0 \text{ for all } \lambda \geq 0 \text{ and } s \in \partial f(x)\} = [\mathbb{R}^+ \partial f(x)]^\circ. \end{aligned}$$

The result follows by taking the polar cone of both sides, and observing that, by assumption,  $\mathbb{R}^+ \partial f(x)$  is closed (Proposition III.1.4.7):

$$N_{Sf(x)}(x) = \text{cl}[\mathbb{R}^+ \partial f(x)] = \mathbb{R}^+ \partial f(x). \quad \square$$

**Remark 1.3.6** The assumption  $0 \notin \partial f(x)$ , required by the above two results, can be formulated in a number of equivalent ways:

- In view of Definition 1.1.4, it means  $f'(x, d_0) < 0$  for some  $d_0$ .
- Using the other definition (1.2.1), there is some  $x_0$  such that  $f(x_0) < f(x)$ .
- The latter implies that the same assumption holds everywhere on the level-set  $f(\cdot) = f(x)$ .

As a result, the existence of one point  $x$  with  $0 \notin \partial f(x)$  allows the computation of the tangent and normal cone to the corresponding sublevel-set  $Sf(x)$  at all its points: on its boundary – which, thanks to (1.3.6), is the level-set  $f(\cdot) = f(x)$  – and on its interior (trivial case).  $\square$

Figure 1.3.2 illustrates these results. It is similar to Fig. 1.3.1, except that it is drawn in  $\mathbb{R}^n$ . Its left part represents the horizontal cut of Fig. 1.3.1 at level  $f(x)$ , i.e.

$$T_{\text{epi } f}(x, f(x)) \cap \{(d, r) \in \mathbb{R}^n \times \mathbb{R} : r = 0\}$$

which is  $T_{Sf(x)}(x) \times \{0\}$ ; considered as a set in  $\mathbb{R}^n$ , it is the tangent cone to the sublevel-set  $Sf(x)$ . In the right part of the picture, we have also drawn the subdifferential, neglecting its vertical component drawn in Fig. 1.3.1. The cone  $N_{Sf(x)}(x)$  generated by this subdifferential appears to be the projection (onto the same horizontal space) of the normal cone  $N_{\text{epi } f}(x, f(x))$ .

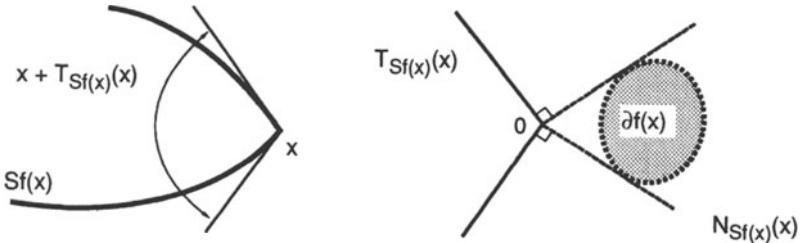


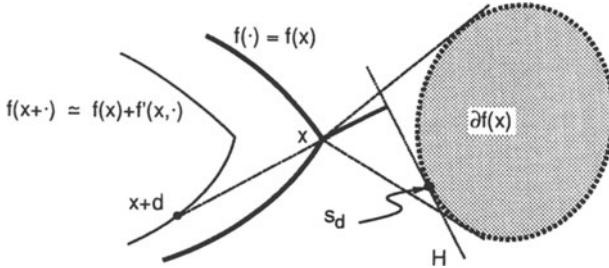
Fig. 1.3.2. Tangent and normal cones to a sublevel-set

All this confirms how a subdifferential generalizes a gradient: if  $\partial f(x)$  is the singleton  $\{\nabla f(x)\}$ , the level-set  $f(\cdot) = f(x)$  has a tangent hyperplane at  $(x, f(x))$ ; the cone  $T_{Sf(x)}(x)$  is the half-space opposite to  $\nabla f(x)$ ; the cone  $N_{Sf(x)}(x)$  is the half-line  $\mathbb{R}^+ \nabla f(x)$ . When  $\partial f(x)$  becomes “fatter”,  $Sf(x)$  becomes “narrower” around  $x$ .

Drawing the normal cone to the sublevel-set, i.e. considering only nonnegative multiples of the subgradients, does describe the sublevel-set locally around  $x$ ; but it also destroys some information, namely the magnitudes of these gradients. This information contains the magnitudes of the directional derivatives, and Fig. 1.3.3 shows how to recover it, even in  $\mathbb{R}^n$ . It is similar to Fig. 1.3.2 and should be compared to Fig. V.2.1.1: the supporting hyperplane

$$H := \{s \in \mathbb{R}^n : \langle d, s - x \rangle = f'(x, d)\}$$

is orthogonal to  $d$ ; its (algebraic) distance to  $x$  is  $f'(x, d)$ , if  $d$  is normalized. The half-line  $x + \mathbb{R}^+ d$  would cut the sublevel-set  $S_{f(x)-1}(f)$  at the (algebraic) distance  $f'(x, d)$  if  $t \mapsto f(x + td)$  were an affine function of  $t \geq 0$ .



**Fig. 1.3.3.** Rate of decrease along a direction

## 1.4 A Constructive Approach to the Existence of a Subgradient

The existence of a subgradient of  $f$  at  $x$  results from Lemma V.3.1.1:  $\partial f(x)$  of Definition 1.1.4 is nonempty just because the closed sublinear function  $f'(x, \cdot)$  supports a nonempty set. Alternatively, such a subgradient can be singled out in Definition 1.2.1 because  $f$  is minorized by some affine function (Proposition IV.1.2.1). In both cases, the seminal argument is the separation Theorem III.4.1.1.

We mention here an interesting alternative construction of a linear function minorizing a (finite) sublinear function  $\sigma$  – standing for  $f'(x, \cdot)$ . The key idea will be as follows: take the directional derivative of  $\sigma$  at some  $d \neq 0$ . It is easy to realize from positive homogeneity that

$$\sigma'(d, d) + \sigma'(d, -d) = 0.$$

In other words, the sublinear function  $\sigma'(d, \cdot)$  is linear at least on the 1-dimensional subspace generated by  $d$  (remember Theorem V.1.1.6). Furthermore, the following result ensures that the subspace where  $\sigma$  was already linear is not spoiled when passing to  $\sigma'(d, \cdot)$ .

**Lemma 1.4.1** *Let a subspace  $U \subset \mathbb{R}^n$  and two functions  $\sigma_1$  and  $\sigma_2$  satisfy the following properties:  $\sigma_1$  is linear on  $U$ ,  $\sigma_2$  is sublinear on  $U$ , and*

$$\sigma_2(x) \leq \sigma_1(x) \quad \text{for all } x \in U.$$

*Then there actually holds*

$$\sigma_2(x) = \sigma_1(x) \quad \text{for all } x \in U.$$

**PROOF.** For all  $x \in U$ , we have

$$0 \leq \sigma_2(x) + \sigma_2(-x) \leq \sigma_1(x) + \sigma_1(-x) = 0,$$

so we conclude

$$\sigma_2(x) = -\sigma_2(-x) \geq -\sigma_1(-x) = \sigma_1(x).$$

□

Then, given our sublinear function  $\sigma$ , let  $\{e_1, \dots, e_n\}$  be a basis of  $\mathbb{R}^n$  and define recursively the following sublinear functions:

$$\sigma_0 := \sigma \quad \text{and} \quad \sigma_k := \sigma'_{k-1}(e_k, \cdot) \text{ for } k = 1, \dots, n. \quad (1.4.1)$$

By now, it should be clear that the subspace on which  $\sigma_k$  is linear increases by at least one more dimension at each  $k$ : we must end up with a linear function.

**Theorem 1.4.2** *In the process (1.4.1), we have*

$$\sigma_n \leq \dots \leq \sigma_k \leq \dots \leq \sigma_0 = \sigma. \quad (1.4.2)$$

*It follows that  $\sigma_n$  is linear and minorizes  $\sigma$ . Moreover,  $\sigma_n(e_1) = \sigma(e_1)$ .*

PROOF. From the definition of  $\sigma_k$  and sublinearity:

$$\sigma_k(d) = \inf_{t>0} \frac{\sigma_{k-1}(e_k + td) - \sigma_{k-1}(e_k)}{t} \leq \sigma_{k-1}(d) \quad \text{for all } d \in \mathbb{R}^n,$$

which proves (1.4.2).

Now, by definition of  $\sigma_k$ , for  $k = 1, \dots, n$

$$\sigma_k(e_k) = \lim_{t \downarrow 0} \frac{(1+t)\sigma_{k-1}(e_k) - \sigma_{k-1}(e_k)}{t} = \sigma_{k-1}(e_k) \quad (1.4.3)$$

where we have used positive homogeneity of  $\sigma_{k-1}$ ; likewise

$$\sigma_k(-e_k) = \lim_{t \downarrow 0} \frac{(1-t)\sigma_{k-1}(e_k) - \sigma_{k-1}(e_k)}{t} = -\sigma_{k-1}(e_k).$$

We deduce simply by addition

$$\sigma_k(e_k) + \sigma_k(-e_k) = \sigma_{k-1}(e_k) - \sigma_{k-1}(e_k) = 0 \quad \text{for } k = 1, \dots, n.$$

In view of Theorem V.1.1.6, each  $\sigma_k$  is linear on the 1-dimensional subspace generated by  $e_k$ . Recursively, Lemma 1.4.1 together with (1.4.2) implies that each  $\sigma_k$  is linear on the subspace generated by  $\{e_j\}_{j \leq k}$ :  $\sigma_n$  is linear on the whole space.

Finally, observe from (1.4.3) that  $\sigma_1(e_1) = \sigma(e_1)$ . Since  $\sigma_1$  is linear on the subspace generated by  $e_1$ , Lemma 1.4.1 again implies recursively

$$\sigma(e_1) = \sigma_1(e_1) = \dots = \sigma_n(e_1). \quad \square$$

In summary, any sublinear function such as  $f'(x, \cdot)$  is minorized by a linear function  $\ell$ . This is essentially the so-called Hahn-Banach Theorem in analytical form; indeed,  $\ell$  supports a singleton  $\{s\}$ , which is a subgradient of  $f$  at  $x$ , and there holds for all  $y \in \mathbb{R}^n$

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

i.e.  $\ell$  discloses a hyperplane supporting the convex set  $\text{epi } f$  at  $(x, f(x))$  (Hahn-Banach Theorem in geometric form).

The following example shows how the process (1.4.1) works in practice.

**Example 1.4.3** For  $d = (\delta^1, \dots, \delta^n)$ , suppose that our initial sublinear function is

$$[f'(x, d) =] \sigma_0(d) := \max \{\delta^1, \dots, \delta^n\},$$

and take  $e_1 := (1, 1, \dots, 1)$ ,  $e_2 := (0, 1, \dots, 1), \dots, e_n := (0, \dots, 0, 1)$  forming a basis of  $\mathbb{R}^n$ . It is not too difficult to see that

$$\sigma_k(d) = \lim_{t \downarrow 0} \frac{\max \{t\delta^1, \dots, t\delta^{k-1}, 1+t\delta^k, \dots, 1+t\delta^n\} - 1}{t} = \max \{\delta^k, \dots, \delta^n\}$$

so  $\sigma_n = \langle e_n, \cdot \rangle$  (we take the standard dot-product for  $\langle \cdot, \cdot \rangle$ ).

This example is interesting because no  $\sigma_k$  is linear for  $k < n$ : the process does take  $n$  steps, although  $\sigma_0$  is already linear along  $e_1$ . The reason is that  $\sigma_1 = \sigma_0$ : the first step is useless. Figure 6.3.2 will clearly show that

- if iterated in the order  $e_2, \dots, e_n, e_1$ , the process takes  $n - 1$  steps:  $\sigma_{n-1}$  is linear;
- if started on  $e_n$ , the linear  $\langle e_n, \cdot \rangle$  is already produced at the first iteration.  $\square$

## 2 Local Properties of the Subdifferential

In this section, we study some properties of  $\partial f(x)$ , considered as a generalization of the concept of gradient, at a given fixed  $x$ .

### 2.1 First-Order Developments

As already mentioned, a finite convex function enjoys a “directional first-order approximation” (1.1.7), and an important result is that the convergence in (1.1.7) is *uniform* in  $d$  on any bounded set:  $\varepsilon_d$  can be taken independent of the normalized  $d$ .

**Lemma 2.1.1** *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be convex and  $x \in \mathbb{R}^n$ . For any  $\varepsilon > 0$ , there exists  $\delta > 0$  such that  $\|h\| \leq \delta$  implies*

$$|f(x + h) - f(x) - f'(x, h)| \leq \varepsilon \|h\|. \quad (2.1.1)$$

PROOF. Suppose for contradiction that there is  $\varepsilon > 0$  and a sequence  $\{h_k\}$  with  $\|h_k\| =: t_k \leq 1/k$  such that

$$|f(x + h_k) - f(x) - f'(x, h_k)| > \varepsilon t_k \quad \text{for } k = 1, 2, \dots$$

Extracting a subsequence if necessary, assume that  $h_k/t_k \rightarrow d$  for some  $d$  of norm 1. Then take a local Lipschitz constant  $L$  of  $f$  (see Remark 1.1.3) and expand:

$$\begin{aligned} \varepsilon t_k &< |f(x + h_k) - f(x) - f'(x, h_k)| \\ &\leq |f(x + h_k) - f(x + t_k d)| + \\ &\quad + |f(x + t_k d) - f(x) - f'(x, t_k d)| + |f'(x, t_k d) - f'(x, h_k)| \\ &\leq 2L\|h_k - t_k d\| + |f(x + t_k d) - f(x) - t_k f'(x, d)|. \end{aligned}$$

Divide by  $t_k > 0$  and pass to the limit to obtain the contradiction  $\varepsilon \leq 0$ .  $\square$

Another way of writing (2.1.1) is the first-order expansion

$$f(x + h) = f(x) + f'(x, h) + o(\|h\|), \quad (2.1.2)$$

or also

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

**Remark 2.1.2** Convexity plays a little role for Lemma 2.1.1. Apart from the existence of a directional derivative, the proof uses only Lipschitz properties of  $f$  and  $f'(x, \cdot)$ .

This remark is of general interest. When defining a concept of “derivative”  $D : \mathbb{R}^n \rightarrow \mathbb{R}$  attached to some function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  at  $x \in \mathbb{R}^n$ , one considers the property

$$\frac{f(x + h) - f(x) - D(h)}{\|h\|} \rightarrow 0, \quad (2.1.3)$$

with  $h$  tending to 0. Even without specifying the class of functions that  $D$  belongs to, several sorts of derivatives can be defined, depending on the type of convergence allowed for  $h$ :

- (i) the point of view of Gâteaux: (2.1.3) holds for  $h = td$ ,  $d$  fixed in  $\mathbb{R}^n$ ,  $t \rightarrow 0$  in  $\mathbb{R}$ ;
- (ii) the point of view of Fréchet: (2.1.3) holds for  $h \rightarrow 0$ ;
- (i') the directional point of view: as in (i), but with  $t > 0$ ;
- (ii') the directional point of view of Dini:  $h = td'$ , with  $t \downarrow 0$  and  $d' \rightarrow d$  in  $\mathbb{R}^n$ .

It should be clear from the proof of Lemma 2.1.1 that, once the approximating function  $D$  is specified, these four types of convergence are equivalent when  $f$  is Lipschitzian around  $x$ .  $\square$

Compare (2.1.2) with the radial development (1.1.7), which can be written with any subgradient  $s_d$  maximizing  $\langle s, d \rangle$ . Such an  $s_d$  is an arbitrary element in the face of  $\partial f(x)$  exposed by  $d$  (remember Definition V.3.1.3). Equivalently,  $d$  lies in the normal cone to  $\partial f(x)$  at  $s_d$ ; or also (Proposition III.5.3.3),  $s_d$  is the projection of  $s_d + d$  onto  $\partial f(x)$ . Thus, the following is just a restatement of Lemma 2.1.1.

**Corollary 2.1.3** Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be convex. At any  $x$ ,

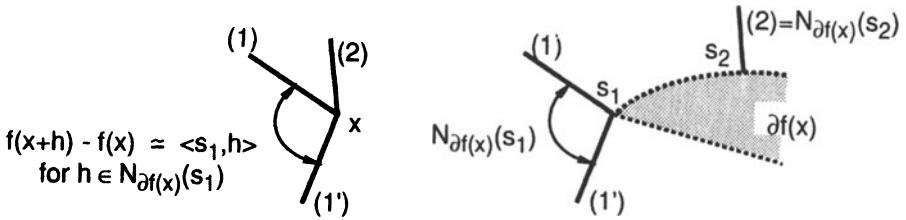
$$f(x + h) = f(x) + \langle s, h \rangle + o(\|h\|)$$

whenever one of the following equivalent properties holds:

$$s \in F_{\partial f(x)}(h) \iff h \in N_{\partial f(x)}(s) \iff s = p_{\partial f(x)}(s + h). \quad \square$$

In words: as long as the increment  $h$  varies in a portion of  $\mathbb{R}^n$  that is in some fixed normal cone to  $\partial f(x)$ ,  $f$  looks differentiable; any subgradient in the corresponding exposed face can be considered as a “local gradient”, active only on that cone; see Fig. 2.1.1. When  $h$  moves to another normal cone, it is another “local gradient” that prevails.

Because  $\partial f(x)$  is compact, any nonzero  $h \in \mathbb{R}^n$  exposes a nonempty face. When  $h$  describes  $\mathbb{R}^n \setminus \{0\}$ , the corresponding exposed faces cover the boundary of  $\partial f(x)$ :



**Fig. 2.1.1.** Apparent differentiability in normal cones

this is Proposition V.3.1.5. An important special case is of course that of  $\partial f(x)$  with only one exposed face, i.e. only one element. This means that there is some *fixed*  $s \in \mathbb{R}^n$  such that

$$\lim_{t \downarrow 0} \frac{f(x + td) - f(x)}{t} = \langle s, d \rangle \quad \text{for all } d \in \mathbb{R}^n,$$

which expresses precisely the *Gâteaux differentiability* of  $f$  at  $x$ . From Corollary 2.1.3 (see again Remark 2.1.2), this is further equivalent to

$$f(x + h) - f(x) = \langle s, h \rangle + o(\|h\|) \quad \text{for all } h \in \mathbb{R}^n,$$

i.e.  $f$  is *Fréchet differentiable* at  $x$ . We will simply say that our function is differentiable at  $x$ , a non-ambiguous terminology.

We summarize our observations:

**Corollary 2.1.4** *If the convex  $f$  is (Gâteaux) differentiable at  $x$ , its only subgradient at  $x$  is its gradient  $\nabla f(x)$ . Conversely, if  $\partial f(x)$  contains only one element  $s$ , then  $f$  is (Fréchet) differentiable at  $x$ , with  $\nabla f(x) = s$ .*  $\square$

Note the following consequence of Proposition V.1.1.6: if  $\{d_1, \dots, d_k\}$  is a set of vectors generating the whole space and if  $f'(x, d_i) = -f'(x, -d_i)$  for  $i = 1, \dots, k$ , then  $f$  is differentiable at  $x$ . In particular (take  $\{d_i\}$  as the canonical basis of  $\mathbb{R}^n$ ), the *existence alone* of the partial derivatives

$$\frac{\partial f}{\partial \xi^i}(x) = f'(x, e_i) = -f'(x, -e_i) \quad \text{for } i = 1, \dots, n$$

guarantees the differentiability of the convex  $f$  at  $x = (\xi^1, \dots, \xi^n)$ . See again Proposition IV.4.2.1.

For the general case where  $\partial f(x)$  is not a singleton, we mention here another way of defining faces: the function  $f'(x, \cdot)$  being convex, it has subdifferentials in its own right (Proposition 1.1.6 studied the subdifferential at 0 only). These subdifferentials are precisely the exposed faces of  $\partial f(x)$ .

**Proposition 2.1.5** *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be convex. For all  $x$  and  $d$  in  $\mathbb{R}^n$ , we have*

$$F_{\partial f(x)}(d) = \partial[f'(x, \cdot)](d).$$

PROOF. If  $s \in \partial f(x)$  then

$$f'(x, d') \geq \langle s, d' \rangle \quad \text{for all } d' \in \mathbb{R}^n$$

simply because  $f'(x, \cdot)$  is the support function of  $\partial f(x)$ . If, in addition,  $\langle s, d \rangle = f'(x, d)$ , we get

$$f'(x, d') \geq f'(x, d) + \langle s, d' - d \rangle \quad \text{for all } d' \in \mathbb{R}^n \quad (2.1.4)$$

which proves the inclusion  $F_{\partial f(x)}(d) \subset \partial[f'(x, \cdot)](d)$ .

Conversely, let  $s$  satisfy (2.1.4). Set  $d'' := d' - d$  and deduce from subadditivity

$$f'(x, d) + f'(x, d'') \geq f'(x, d') \geq f'(x, d) + \langle s, d'' \rangle \quad \text{for all } d'' \in \mathbb{R}^n$$

which implies  $f'(x, \cdot) \geq \langle s, \cdot \rangle$ , hence  $s \in \partial f(x)$ . Also, putting  $d' = 0$  in (2.1.4) shows that  $\langle s, d \rangle \geq f'(x, d)$ . Altogether, we have  $s \in F_{\partial f(x)}(d)$ .  $\square$

This result is illustrated in Fig. 2.1.2. Observe in particular that the subdifferential of  $f'(x, \cdot)$  at the point  $td$  does not depend on  $t > 0$ ; but when  $t$  reaches 0, this subdifferential explodes to the entire  $\partial f(x)$ .

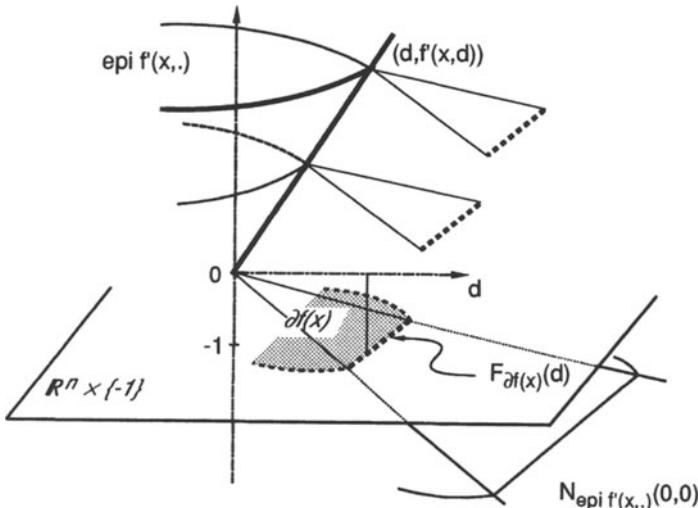


Fig. 2.1.2. Faces of subdifferentials

**Definition 2.1.6** A point  $x$  at which  $\partial f(x)$  has more than one element – i.e. at which  $f$  is not differentiable – is called a *kink* (or corner-point) of  $f$ .  $\square$

We know that  $f$  is differentiable almost everywhere (Theorem IV.4.2.3). The set of kinks is therefore of zero-measure. In most examples in practice, this set is the union of a finite number of algebraic surfaces in  $\mathbb{R}^n$ .

**Example 2.1.7** Let  $q_1, \dots, q_m$  be  $m$  convex quadratic functions and take  $f$  as the max of the  $q_j$ 's. Given  $x \in \mathbb{R}^n$ , let

$$J(x) := \{j \leq m : q_j(x) = f(x)\}$$

denote the set of active  $q_j$ 's at  $x$ .

It is clear that  $f$  is differentiable at each  $x$  such that  $J(x)$  reduces to a singleton  $\{j(x)\}$ : by continuity,  $J(y)$  is still this singleton  $\{j(x)\}$  for all  $y$  close enough to  $x$ , so  $f$  and its gradient coincide around this  $x$  with the smooth  $q_{j(x)}$  and its gradient. Thus, our  $f$  has all its kinks in the union of the  $1/2m(m - 1)$  surfaces

$$\Sigma_{ij} := \{x \in \mathbb{R}^n : q_i(x) = q_j(x)\} \quad \text{for } i \neq j.$$

Figure 2.1.3 gives an idea of what this case could look like, in  $\mathbb{R}^2$ . The dashed lines represent portions of  $\Sigma_{ij}$  at which  $q_i = q_j < f$ .  $\square$

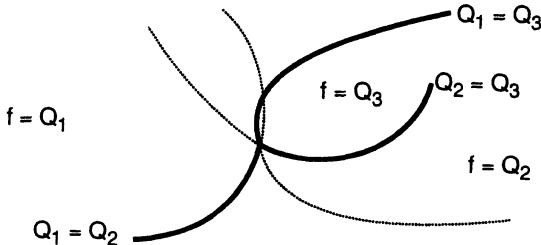


Fig. 2.1.3. A maximum of convex quadratic functions

## 2.2 Minimality Conditions

We start with a fundamental result, coming directly from the definitions of the subdifferential.

**Theorem 2.2.1** For  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  convex, the following three properties are equivalent:

- (i)  $f$  is minimized at  $x$  over  $\mathbb{R}^n$ , i.e.:  $f(y) \geq f(x)$  for all  $y \in \mathbb{R}^n$ ;
- (ii)  $0 \in \partial f(x)$ ;
- (iii)  $f'(x, d) \geq 0$  for all  $d \in \mathbb{R}^n$ .

PROOF. The equivalence (i)  $\Leftrightarrow$  (ii) [resp. (ii)  $\Leftrightarrow$  (iii)] is obvious from (1.2.1) [resp. (1.1.6)].  $\square$

Naturally,  $x$  can be called “stationary” if  $0 \in \partial f(x)$ . Observe that the equivalence (i)  $\Leftrightarrow$  (iii) says:  $f$  is minimal at  $x$  if and only if its tangential approximation  $f'(x, \cdot)$  is minimal at 0; a statement which makes sense, and which calls for two remarks.

- When  $x$  is a local minimum of  $f$  (in the sense of Definition II.1.1.2), (iii) holds; thus, convexity implies that a local minimum is automatically global.

- In the smooth case, the corresponding statement would be “the tangential approximation  $\langle \nabla f(x), \cdot \rangle$  of  $f(\cdot) - f(x)$ , which is linear, is identically 0”. Here, the tangential approximation  $f'(x, \cdot)$  need not be 0; but there does exist a minorizing linear function, usually not tangential, which is identically 0.

**Remark 2.2.2** The property “ $0 \in \partial f(x)$ ” is a generalization of the usual stationarity condition “ $\nabla f(x) = 0$ ” of the smooth case. Even though the gradient exists at almost every  $x$ , one should not think that a convex function has almost certainly a 0-gradient at a minimum point. As a matter of fact, the “probability” that a given  $x$  is a kink is 0; but this probability may not stay 0 if some more information is known about  $x$ , for example that it is a stationary point. As a rule, the minimum points of a convex function are indeed kinks.  $\square$

The position of 0 in  $\partial f(x)$  is useful for determining the nature of  $x$  as a minimum point: it tells us how  $f$  increases in the neighborhood of  $x$ . In the smooth case,  $f$  is stationary at  $x$  if and only if  $\partial f(x)$  is the singleton  $\{0\}$ . This means that the first-order approximation of  $f$  at  $x$  is constant. In addition, convexity implies that  $f$  is really minimal at  $x$  – and not maximal, say. If we ask more about the behaviour of  $f$  around  $x$ , not much can be extracted from the property “ $\nabla f(x) = 0$ ” alone. It cannot be ascertained, for example, whether  $x$  is a unique minimum. In the nonsmooth case, the possible existence of additional nonzero subgradients makes the geometry of the graph of  $f$  much more versatile. The essential result is the following.

**Proposition 2.2.3** *Let  $x$  minimize the convex  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  and let  $N_{\partial f(x)}(0)$  denote the normal cone to  $\partial f(x)$  at 0.*

*For all  $\varepsilon > 0$  there exists  $\delta > 0$  such that*

$$h \in N_{\partial f(x)}(0) \cap B(0, \delta) \implies f(x + h) \leq f(x) + \varepsilon \|h\|. \quad (2.2.1)$$

*On the other hand,*

$$h \notin N_{\partial f(x)}(0) \implies f(x + h) > f(x). \quad (2.2.2)$$

PROOF. By definition,  $h \in N_{\partial f(x)}(0)$  if and only if

$$\langle s, h \rangle \leq 0 \quad \text{for all } s \in \partial f(x) \quad (2.2.3)$$

and, knowing that  $x$  minimizes  $f$ , this is equivalent to  $f'(x, h) = 0$ . Then (2.2.1) is a consequence of the first-order development (2.1.2).

When  $h \notin N_{\partial f(x)}(0)$ , the negation of (2.2.3) is  $f'(x, h) > 0$ , whence (2.2.2) follows immediately.  $\square$

**Remark 2.2.4** A somewhat more accurate statement than (2.2.2) can be given, taking into account the *direction* of  $h \neq 0$ : for all  $d \notin N_{\partial f(x)}(0)$ , there is  $\varepsilon > 0$  such that

$$f(x + td) \geq f(x) + \varepsilon t \quad \text{for all } t \geq 0. \quad (2.2.4)$$

Of course,  $\varepsilon$  certainly depends on  $d$ : it is nothing but  $f'(x, d)$ , which is positive since  $d \notin N_{\partial f(x)}(0)$ .  $\square$

Proposition 2.2.3 is important for optimization theory, because it divides the space around a minimum point  $x$  in two parts:

- The first part is the normal cone  $N_{\partial f(x)}(0)$  to  $\partial f(x)$  at 0. For a direction  $d$  in this normal cone,  $f'(x, d) = 0$  and the function  $\varphi : 0 \leq t \mapsto f(x + td)$  is constant to first order at  $t = 0$ . For such a  $d$ , it might be the case, for example, that  $\varphi(t) = \varphi(0)$  for small  $t \geq 0$ . To check whether or not  $\varphi$  is strictly increasing, and at what speed it does so, requires more work; the situation is similar in the smooth case, where a second-order analysis is required.
- When  $d$  is out of this normal cone,  $f(x + td)$  increases with a nonzero linear rate: if only this part of the space were concerned,  $x$  would be a guaranteed unique minimum.

**Remark 2.2.5** The normal cone  $N_{\partial f(x)}(0)$  can thus be called the *critical cone*, its nonzero elements being the *critical directions*. In classical (smooth) optimization, the space around a local minimum  $x$  is divided analogously into two regions:

- One is a subspace: the kernel of  $\nabla^2 f(x)$ , sometimes called the set of critical directions. For all  $\varepsilon > 0$ , there exists  $\delta > 0$  such that

$$h \in \text{Ker } \nabla^2 f(x) \cap B(0, \delta) \implies f(x + h) \leq f(x) + \varepsilon \|h\|^2. \quad (2.2.5)$$

- In the complement of this subspace,  $f$  increases as fast as a strongly convex quadratic function: for all  $d \notin \text{Ker } \nabla^2 f(x)$ , there is  $\varepsilon > 0$  such that

$$f(x + td) \geq f(x) + \varepsilon t^2 \quad \text{for } t \text{ close enough to } 0. \quad (2.2.6)$$

The present nonsmooth situation is fairly similar, if we replace subspaces by cones and  $t^2$  by  $t$ : compare (2.2.5) with (2.2.1) on the one hand, (2.2.6) with (2.2.4) on the other. A substantial difference, however, is that the existence of critical directions is now the rule; by contrast, in smooth optimization, the assumption  $\text{Ker } \nabla^2 f(x) = \{0\}$  – i.e.  $\nabla^2 f(x)$  is positive definite – is well-accepted.

As mentioned earlier, the possible property  $N_{\partial f(x)}(0) \neq \mathbb{R}^n$  is a privilege of nonsmooth functions. It brings some definite advantages, one being that a first-order analysis may sometimes suffice to guarantee uniqueness of a minimum point.  $\square$

There are various interesting special cases of Proposition 2.2.3. For example, the property “ $0 \in \text{ri } \partial f(x)$ ” is equivalent to

$$f'(x, d) > 0 \quad \text{for all } d \text{ with } f'(x, d) + f'(x, -d) > 0$$

(remember Theorem V.2.2.3). In the language of Remark 1.1.5, this last property means that  $x$  is a strict minimum in all directions along which  $f$  is not smooth. A “super-special” case arises when, in addition,  $\partial f(x)$  is full-dimensional, i.e.

$$0 \in \text{ri } \partial f(x) = \text{int } \partial f(x) \neq \emptyset.$$

Then there are no critical directions.

**Proposition 2.2.6** *A necessary and sufficient condition for the existence of  $\varepsilon > 0$  such that*

$$f(x + h) \geq f(x) + \varepsilon \|h\| \quad \text{for all } h \in \mathbb{R}^n \quad (2.2.7)$$

*is  $0 \in \text{int } \partial f(x)$ .*

PROOF. The condition  $0 \in \text{int } \partial f(x)$  means that  $B(0, \varepsilon) \subset \partial f(x)$  for some  $\varepsilon > 0$  which, in terms of support functions, can be written as  $f'(x, \cdot) \geq \varepsilon \|\cdot\|$ . From the definition of the directional derivative, this last property is equivalent to (2.2.7).  $\square$

With the above results in mind, it is instructive to look once more at Fig. 1.1.1. Take  $x$  to be a minimum point (on the right part of the picture, translate  $\partial f(x)$  down to  $V$ ). Then the critical cone is just the subspace  $U$ , in which  $f'(x, \cdot)$  is linear – and identically zero.

The case illustrated by this figure is rather typical: the subdifferential at a minimum point is often not full-dimensional; it has therefore an empty interior, and Proposition 2.2.6 does not apply. In this case, the critical cone is definitely nontrivial: it is  $U$ .

On the other hand, this cone often coincides with this subspace, which means that  $0 \in \text{ri } \partial f(x)$ . Still in the same Fig. 1.1.1, translate  $\partial f(x)$  further to the right so as to place 0 at its left endpoint. Then the critical cone becomes the left half-space. This is not a typical situation.

Let us sum up:

- A minimum point  $x$  is characterized by:  $0 \in \partial f(x)$ , or  $f'(x, d) \geq 0$  for all  $d \in \mathbb{R}^n$ .
- A *critical direction* is a  $d \neq 0$  such that  $f'(x, d) = 0$ . Existence of a critical direction is equivalent to  $0 \in \text{bd } \partial f(x)$ .
- If there is a critical  $d$  with  $-d$  non-critical,  $x$  can be considered as degenerate. This is equivalent to  $0 \in \text{rbd } \partial f(x)$ .

To finish, we mention that a non-minimal  $x$  is characterized by the existence of a  $d$  with  $f'(x, d) < 0$ . Such a  $d$  is called a *descent direction*, a concept which plays an important role for minimization algorithms. The set of descent directions was shown in Theorem 1.3.4 to be the interior of the tangent cone to  $Sf(x)$ , the sublevel-set passing at  $x$ .

### 2.3 Mean-Value Theorems

Given two distinct points  $x$  and  $y$ , and knowing the subdifferential of  $f$  on the whole line-segment  $[x, y]$ , can we evaluate  $f(y) - f(x)$ ? Or also, is it possible to express  $f$  as the integral of its subdifferential? This is the aim of mean-value theorems.

Of course, the problem reduces to that of one-dimensional convex functions (Chap. I), since

$$f(y) - f(x) = \varphi(1) - \varphi(0)$$

where

$$\varphi(t) := f(ty + (1-t)x) \quad \text{for all } t \in [0, 1] \quad (2.3.1)$$

is the trace of  $f$  on the line-segment  $[x, y]$ . The key question, however, is to express the subdifferential of  $\varphi$  at  $t$  in terms of the subdifferential of  $f$  at  $ty + (1-t)x$  in the surrounding space  $\mathbb{R}^n$ . The next lemma anticipates the calculus rules to be given in §4. Here and below, we use the following notation:

$$x_t := ty + (1-t)x$$

where  $x$  and  $y$  are considered as fixed in  $\mathbb{R}^n$ .

**Lemma 2.3.1** *The subdifferential of  $\varphi$  defined by (2.3.1) is*

$$\partial\varphi(t) = \{\langle s, y - x \rangle : s \in \partial f(x_t)\}$$

or, more symbolically:

$$\partial\varphi(t) = (\partial f(x_t), y - x).$$

PROOF. Apply the definitions from §I.4:

$$D_+\varphi(t) = \lim_{\tau \downarrow 0} \frac{f(x_t + \tau(y - x)) - f(x_t)}{\tau} = f'(x_t, y - x)$$

$$D_-\varphi(t) = \lim_{\tau \uparrow 0} \frac{f(x_t + \tau(y - x)) - f(x_t)}{\tau} = -f'(x_t, -(y - x))$$

so, knowing that

$$f'(x_t, y - x) = \max_{s \in \partial f(x_t)} \langle s, y - x \rangle,$$

$$-f'(x_t, -(y - x)) = \min_{s \in \partial f(x_t)} \langle s, y - x \rangle,$$

we obtain

$$\partial\varphi(t) := [D_-\varphi(t), D_+\varphi(t)] = \{\langle s, y - x \rangle : s \in \partial f(x)\}. \quad \square$$

**Remark 2.3.2** The derivative of  $\varphi$  exists except possibly on a countable set in  $\mathbb{R}$ . One should not think that, with this pretext,  $f$  is differentiable except possibly at countably many points of  $[x, y]$ . For example, with  $f(\xi, \eta) := |\xi|$ ,  $x := (0, 0)$ ,  $y := (0, 1)$ ,  $f$  is differentiable nowhere on  $[x, y]$ . What Lemma 2.3.1 ensures, however, is that for almost all  $t$ ,  $\partial f(x_t)$  has a zero breadth in the direction  $y - x$ :  $f'(x_t, y - x) + f'(x_t, x - y) = 0$ .

Note, on the other hand – and this is a consequence of Fubini’s theorem – that almost all the lines parallel to  $[x, y]$  have an intersection of zero-measure with the set of kinks of  $f$ .  $\square$

With the calculus rule given in Lemma 2.3.1, the one-dimensional mean-value Theorem I.4.2.4 applied to the function  $\varphi$  of (2.3.1) becomes the following result:

**Theorem 2.3.3** *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be convex. Given two points  $x \neq y$  in  $\mathbb{R}^n$ , there exist  $t \in ]0, 1[$  and  $s \in \partial f(x_t)$  such that*

$$f(y) - f(x) = \langle s, y - x \rangle. \quad (2.3.2)$$

In other words,

$$f(y) - f(x) \in \bigcup_{t \in ]0, 1[} \{\langle \partial f(x_t), y - x \rangle\}. \quad \square$$

The mean-value theorem can also be given in integral form.

**Theorem 2.3.4** Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be convex. For  $x, y \in \mathbb{R}^n$ ,

$$f(y) - f(x) = \int_0^1 \langle \partial f(x_t), y - x \rangle dt. \quad (2.3.3)$$
□

The meaning of (2.3.3) is as follows: if  $\{s_t : t \in [0, 1]\}$  is any selection of subgradients of  $f$  on the line-segment  $[x, y]$ , i.e.  $s_t \in \partial f(x_t)$  for all  $t \in [0, 1]$ , then  $\int_0^1 \langle s_t, y - x \rangle dt$  is independent of the selection and its value is  $f(y) - f(x)$ .

**Example 2.3.5** Mean-value theorems can be applied to nondifferentiable functions in the same way as they are in (ordinary) differential calculus. As an example, let  $f, g$  be two convex functions such that  $f(x) = g(x)$  and  $f(y) \leq g(y)$  for all  $y$  in a neighborhood of  $x$ . It follows from the definitions that  $\partial f(x) \subset \partial g(x)$ .

Conversely, suppose that  $f, g$  are such that  $\partial f(x) \subset \partial g(x)$  for all  $x \in \mathbb{R}^n$ ; can we compare  $f$  and  $g$ ? Same question if  $\partial f(x) \cap \partial g(x) \neq \emptyset$  for all  $x$ . The answer lies in (2.3.3): the difference  $f - g$  is a constant function (take the same selection in the integral (2.3.3)!).

An amusing particular case is that where  $f$  and  $g$  are (finite) sublinear functions: then,  $f \leq g$  if and only if  $\partial f(0) \subset \partial g(0)$  ( $f$  and  $g$  are the support functions of  $\partial f(0)$  and  $\partial g(0)$  respectively!).

□

### 3 First Examples

**Example 3.1 (Support Functions)** Let  $C$  be a nonempty convex compact set, with support function  $\sigma_C$ . The first-order differential elements of  $\sigma_C$  at the origin are obtained immediately from the definitions:

$$\partial\sigma_C(0) = C \quad \text{and} \quad (\sigma_C)'(0, \cdot) = \sigma_C.$$

Read this with Proposition 1.1.6 in mind: any convex compact set  $C$  can be considered as the subdifferential of some finite convex function  $f$  at some point  $x$ . The simplest instance is  $f = \sigma_C, x = 0$ .

On the other hand, the first-order differential elements of a support function  $\sigma_C$  at  $x \neq 0$  are given in Proposition 2.1.5:

$$\partial\sigma_C(x) = F_C(x) \quad \text{and} \quad (\sigma_C)'(x, \cdot) = \sigma_{F_C}(x). \quad (3.1)$$

The expression of  $(\sigma_C)'(x, d)$  above is a bit tricky: it is the optimal value of the following optimization problem ( $s$  is the variable,  $x$  and  $d$  are fixed, the objective function is linear, there is one linear constraint in addition to those describing  $C$ ):

$$\left| \begin{array}{ll} \max \langle d, s \rangle & s \in C \\ \langle s, x \rangle = \sigma_C(x). \end{array} \right.$$

As a particular case, take a norm  $\|\cdot\|$ . As seen already in §V.3.2, it is the gauge of its unit ball  $B$ , and it is the support function of the unit ball  $B^*$  associated with the dual norm  $\|\cdot\|^*$ . Hence

$$\partial\|\cdot\|(0) = B^* = \{s \in \mathbb{R}^n : \max_{\|d\| \leq 1} \langle s, d \rangle \leq 1\}.$$

More generally, for  $x$  not necessarily zero, (3.1) can be rewritten as

$$\partial \|\cdot\|(x) = \{s \in B^* : \langle s, x \rangle = \max_{u \in B^*} \langle u, x \rangle = \|\cdot\|(x)\}. \quad (3.2)$$

All the points  $s$  in (3.2) have dual norm 1; they form the face of  $B^*$  exposed by  $x$ .  $\square$

**Example 3.2 (Gauges)** Suppose now that, rather than being compact, the  $C$  of Example 3.1 is closed and contains the origin as an interior point. Then, another example of a convex finite function is its gauge  $\gamma_C$  (Theorem V1.2.5). Taking into account the correspondence between supports and gauges – see Proposition V3.2.4 and Corollary V3.2.5 – (3.1) can be copied if we replace  $C$  by its polar set

$$C^\circ := \{x : \langle s, x \rangle \leq 1 \text{ for all } s \in C\}.$$

So we obtain

$$\partial \gamma_C(0) = C^\circ, \quad (\gamma_C)'(0, \cdot) = \gamma_C,$$

and of course, Proposition 2.1.5 applied to  $C^\circ$  gives at  $x \neq 0$

$$\partial \gamma_C(x) = F_{C^\circ}(x) \quad \text{and} \quad (\gamma_C)'(x, \cdot) = \sigma_{F_{C^\circ}(x)}.$$

Gauges and support functions of elliptic sets deserve a more detailed study. Given a symmetric positive semi-definite operator  $Q$ , define

$$\mathbb{R}^n \ni x \mapsto f(x) := \sqrt{\langle Qx, x \rangle}$$

which is just the gauge of the sublevel-set  $\{x : f(x) \leq 1\}$ . From elementary calculus,

$$\partial f(x) = \{\nabla f(x)\} = \left\{ \frac{Qx}{f(x)} \right\} \quad \text{for } x \notin \text{Ker } Q$$

while, for  $x \in \text{Ker } Q$ ,  $s \in \partial f(x)$  if and only if, for all  $y \in \mathbb{R}^n$

$$\langle s, y - x \rangle \leq \sqrt{\langle Qy, y \rangle} = \sqrt{\langle Q(y - x), y - x \rangle} = \|Q^{1/2}(y - x)\|.$$

From the Cauchy-Schwarz inequality (remember Example V2.3.4), we see that  $\partial f(x)$  is the image by  $Q^{1/2}$  of the unit ball  $B(0, 1)$ .  $\square$

**Example 3.3 (Distance Functions)** Let again  $C$  be closed and convex. Another finite convex function is the distance to  $C$ :

$$d_C(x) := \min \{\|y - x\| : y \in C\},$$

in which the min is attained at the projection  $p_C(x)$  of  $x$  onto  $C$ . The subdifferential of  $d_C$  is

$$\partial d_C(x) = \begin{cases} N_C(x) \cap B(0, 1) & \text{if } x \in C, \\ \left\{ \frac{x - p_C(x)}{\|x - p_C(x)\|} \right\} & \text{if } x \notin C, \end{cases} \quad (3.3)$$

a formula illustrated by Fig. V2.3.1 when  $C$  is a closed convex cone. The case  $x \notin C$  was already proved in Example IV4.1.6; let us complete the proof. Thus, for  $x \in C$ , let  $s \in \partial d_C(x)$ , i.e.

$$d_C(x') \geq \langle s, x' - x \rangle \quad \text{for all } x' \in \mathbb{R}^n.$$

This implies in particular  $\langle s, x' - x \rangle \leq 0$  for all  $x' \in C$ , hence  $s \in N_C(x)$ ; and taking  $x' = x + s$ , we obtain

$$\|s\|^2 \leq d_C(x+s) \leq \|x+s-x\| = \|s\|.$$

Conversely, let  $s \in N_C(x) \cap B(0, 1)$  and, for all  $x' \in \mathbb{R}^n$ , write

$$\langle s, x' - x \rangle = \langle s, x' - p_C(x') \rangle + \langle s, p_C(x') - x \rangle.$$

The last scalar product is nonpositive because  $s \in N_C(x)$  and, with the Cauchy-Schwarz inequality, the property  $\|s\| \leq 1$  gives

$$\langle s, x' - p_C(x') \rangle \leq \|x' - p_C(x')\| = d_C(x').$$

Altogether,  $s \in \partial d_C(x)$ . Note that the set of kinks of  $d_C$  is exactly the boundary of  $C$ . This proves once more that the boundary of a convex set is of zero-measure.

Consider the closed convex cone  $K := N_C(x)$ , whose polar cone is  $K^\circ = T_C(x)$ . As seen in Chap. V, more particularly in (V.3.3.5), the support function of  $K' = K \cap B(0, 1)$  is the distance to  $T_C(x)$ . From (3.3), we see that

$$d'_C(x, \cdot) = d_{T_C(x)} \quad \text{for all } x \in C.$$

Compare also this formula with Proposition III.5.3.5.  $\square$

**Example 3.4 (Piecewise Affine Functions)** Consider the function

$$\mathbb{R}^n \ni x \mapsto f(x) := \max \{f_j(x) : j = 1, \dots, m\} \quad (3.4)$$

where each  $f_j$  is affine:

$$f_j(x) := r_j + \langle s_j, x \rangle \quad \text{for } j = 1, \dots, m.$$

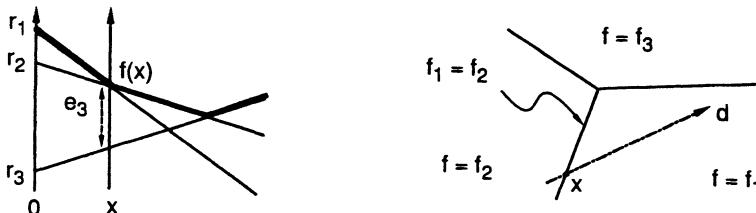
To compute the first-order differential elements of  $f$  at a given  $x$ , it is convenient to translate the origin at this  $x$ . Thus, we rewrite (3.4) as

$$f(y) = f(x) + \max \{-e_j + \langle s_j, y - x \rangle : j = 1, \dots, m\} \quad (3.5)$$

where we have set, for  $j = 1, \dots, m$

$$e_j := f(x) - r_j - \langle s_j, x \rangle = f(x) - f_j(x) \geq 0 \quad (3.6)$$

(look at the left part of Fig. 3.1 to visualize  $e_j$ ).



**Fig. 3.1.** Piecewise affine functions: translation of the origin and directional derivative

Now consider  $f(x+td)$ , as illustrated on the right part of Fig. 3.1, representing the space  $\mathbb{R}^n$  around  $x$ . For  $t > 0$  small enough, those  $j$  such that  $e_j > 0$  do not count. Accordingly, set

$$J(x) := \{j : e_j = 0\} = \{j : f_j(x) = f(x)\}.$$

Rewriting (3.5) again as

$$f(x + td) = f(x) + t \max \{\langle s_j, d \rangle : j \in J(x)\} \quad \text{for small } t > 0,$$

it becomes obvious that

$$f'(x, d) = \max \{\langle s_j, d \rangle : j \in J(x)\}.$$

From the calculus rule V.3.3.3(ii) and Definition 1.1.4, this means exactly that

$$\partial f(x) = \text{co} \{s_j : j \in J(x)\}. \quad (3.7)$$

This result will be confirmed in §4.4. We have demonstrated it here nevertheless in intuitive terms, because piecewise affine functions are of utmost importance. In particular, the notation (3.6) and Fig. 3.1 will be widely used in some of the subsequent chapters, devoted to minimization algorithms.  $\square$

## 4 Calculus Rules with Subdifferentials

Calculus with subdifferentials of convex functions is important for the theory, just as in ordinary differential calculus. Its role is illustrated in Fig. 4.0.1: if  $f$  is constructed from some other convex functions  $f_j$ , the problem is to compute  $\partial f$  in terms of the  $\partial f_j$ 's.

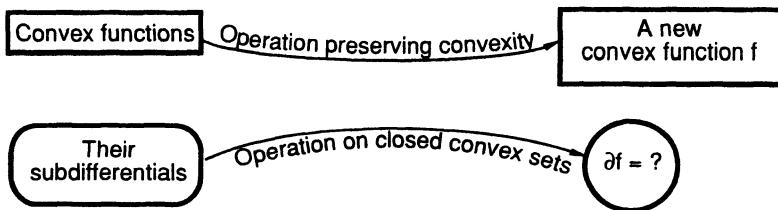


Fig. 4.0.1. Subdifferential calculus

To develop our calculus rules, the two definitions 1.1.4 and 1.2.1 will be used. Calculus with support functions (§V.3.3) will therefore be an essential tool.

### 4.1 Positive Combinations of Functions

**Theorem 4.1.1** *Let  $f_1, f_2$  be two convex functions from  $\mathbb{R}^n$  to  $\mathbb{R}$  and  $t_1, t_2$  be positive. Then*

$$\partial(t_1 f_1 + t_2 f_2)(x) = t_1 \partial f_1(x) + t_2 \partial f_2(x) \quad \text{for all } x \in \mathbb{R}^n. \quad (4.1.1)$$

**PROOF.** Apply Theorem V.3.3.3(i):  $t_1\partial f_1(x) + t_2\partial f_2(x)$  is a compact convex set whose support function is

$$t_1 f'_1(x, \cdot) + t_2 f'_2(x, \cdot). \quad (4.1.2)$$

On the other hand, the support function of  $\partial(t_1 f_1 + t_2 f_2)(x)$  is by definition the directional derivative  $(t_1 f_1 + t_2 f_2)'(x, \cdot)$  which, from elementary calculus, is just (4.1.2). Therefore the two (compact convex) sets in (4.1.1) coincide, since they have the same support function.  $\square$

**Remark 4.1.2** Needless to say, the sign of  $t_1$  and  $t_2$  in (4.1.1) is important to obtain a resulting function which is convex. There is a deeper reason, though: take  $f_1(x) = f_2(x) = \|x\|$ ,  $t_1 = -t_2 = 1$ . We obtain  $f_1 - f_2 \equiv 0$ , yet

$$t_1 \partial f_1(0) + t_2 \partial f_2(0) = B(0, 2),$$

a gross over-estimate of  $\{0\}$ !  $\square$

To illustrate this calculus rule, consider  $f : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$  defined by

$$f(x_1, x_2) = f_1(x_1) + f_2(x_2),$$

with  $f_1$  and  $f_2$  convex on  $\mathbb{R}^p$  and  $\mathbb{R}^q$  respectively. First, call

$$\mathbb{R}^p \times \mathbb{R}^q \ni (x_1, x_2) \mapsto \tilde{f}_1(x_1, x_2) = f_1(x_1)$$

the extension of  $f_1$  and observe that its subdifferential is obviously

$$\partial \tilde{f}_1(x_1, x_2) = \partial f_1(x_1) \times \{0\}.$$

Then Theorem 4.1.1 gives, after the same extension is made with  $f_2$ ,

$$\partial f(x_1, x_2) = \partial f_1(x_1) \times \{0\} + \{0\} \times \partial f_2(x_2) = \partial f_1(x_1) \times \partial f_2(x_2). \quad (4.1.3)$$

**Remark 4.1.3** Given a convex function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , an interesting trick is to view its epigraph as a sublevel-set of a certain convex function, namely:

$$\mathbb{R}^n \times \mathbb{R} \ni (x, r) \mapsto g(x, r) := f(x) - r.$$

Clearly enough,  $\text{epi } f$  is the sublevel-set  $S_0(g)$ . The directional derivatives of  $g$  are easy to compute:

$$g'(x, f(x); d, \rho) = f'(x, d) - \rho \quad \text{for all } (d, \rho) \in \mathbb{R}^n \times \mathbb{R}$$

and (4.1.3) gives for all  $x \in \mathbb{R}^n$

$$\partial g(x, f(x)) = \partial f(x) \times \{-1\} \not\ni 0.$$

We can therefore apply Theorems 1.3.4 and 1.3.5 to  $g$ , which gives back the formulae of Proposition 1.3.1:

$$\begin{aligned} T_{\text{epi } f}(x, f(x)) &= \{(d, \rho) : f'(x, d) \leq \rho\}, \\ \text{int } T_{\text{epi } f}(x, f(x)) &= \{(d, \rho) : f'(x, d) < \rho\} \neq \emptyset, \\ N_{\text{epi } f}(x, f(x)) &= \mathbb{R}^+[\partial f(x) \times \{-1\}]. \end{aligned}$$

$\square$

## 4.2 Pre-Composition with an Affine Mapping

**Theorem 4.2.1** Let  $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$  be an affine mapping ( $Ax = A_0x + b$ , with  $A_0$  linear and  $b \in \mathbb{R}^m$ ) and let  $g$  be a finite convex function on  $\mathbb{R}^m$ . Then

$$\partial(g \circ A)(x) = A_0^* \partial g(Ax) \quad \text{for all } x \in \mathbb{R}^n. \quad (4.2.1)$$

PROOF. Form the difference quotient giving rise to  $(g \circ A)'(x, d)$  and use the relation  $A(x + td) = Ax + tA_0d$  to obtain

$$(g \circ A)'(x, d) = g'(Ax, A_0d) \quad \text{for all } d \in \mathbb{R}^n.$$

From Proposition V.3.3.4, the right-hand side in the above equality is the support function of the convex compact set  $A_0^* \partial g(Ax)$ .  $\square$

This result is illustrated by Lemma 2.3.1: with fixed  $x, y \in \mathbb{R}^n$ , consider the affine mapping  $A : \mathbb{R} \rightarrow \mathbb{R}^n$

$$At := x + t(y - x).$$

Then  $A_0t = t(y - x)$ , and the adjoint of  $A_0$  is defined by

$$A_0^*(s) = \langle y - x, s \rangle \quad \text{for all } s \in \mathbb{R}^n.$$

Twisting the notation, replace  $(n, m, x, g)$  in Theorem 4.2.1 by  $(1, n, t, f)$ : this gives the subdifferential  $\partial\varphi$  of Lemma 2.3.1.

As another illustration, let us come back to the example of §4.1. Needless to say, the validity of (4.1.3) relies crucially on the “decomposed” form of  $f$ . Indeed, take a convex function  $f : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$  and the affine mapping

$$\mathbb{R}^p \ni x_1 \mapsto Ax_1 = (x_1, x_2) \in \mathbb{R}^p \times \mathbb{R}^q.$$

Its linear part is  $x_1 \mapsto A_0x_1 = (x_1, 0)$  and  $A_0^*(s_1, s_2) = s_1$ . Then consider the partial function

$$f \circ A : \mathbb{R}^p \ni x_1 \mapsto f_{x_2}^{(1)}(x_1) = f(x_1, x_2).$$

According to Theorem 4.2.1,

$$\partial f_{x_2}^{(1)}(x_1) = \{s_1 \in \mathbb{R}^p : \exists s_2 \in \mathbb{R}^q \text{ such that } (s_1, s_2) \in \partial f(x_1, x_2)\}$$

is the projection of  $\partial f(x_1, x_2)$  onto  $\mathbb{R}^p$ . Naturally, we can construct likewise the projection of  $\partial f(x_1, x_2)$  onto  $\mathbb{R}^q$ , which yields the inclusion

$$\partial f(x_1, x_2) \subset \partial f_{x_2}^{(1)}(x_1) \times \partial f_{x_1}^{(2)}(x_2). \quad (4.2.2)$$

**Remark 4.2.2** Beware that equality in (4.2.2) need not hold, except in special cases; for example in the decomposable case (4.1.3), or also when one of the projections is a singleton, i.e. when the partial function  $f_{x_2}^{(1)}$ , say, is differentiable. For a counter-example, take  $p = q = 1$  and

$$f(x_1, x_2) = |x_1 - x_2| + \frac{1}{2}(x_1 + 1)^2 + \frac{1}{2}(x_2 + 1)^2.$$

This function has a unique minimum at  $(-1, -1)$  but, at  $(0, 0)$ , we have

$$\partial f_0^{(i)}(0) = [0, 2] \quad \text{for } i = 1, 2,$$

hence the right-hand side of (4.2.2) contains  $(0, 0)$ . Yet,  $f$  is certainly not minimal there,  $\partial f(0, 0)$  is actually the line-segment

$$\{2(\alpha, 1 - \alpha) : \alpha \in [0, 1]\}. \quad \square$$

### 4.3 Post-Composition with an Increasing Convex Function of Several Variables

As seen in §IV.2.1(d), post-composition with an increasing one-dimensional convex function preserves convexity. A relevant object is the subdifferential of the result; we somewhat generalize the problem, by considering a vector-valued version of this operation.

Let  $f_1, \dots, f_m$  be  $m$  convex functions from  $\mathbb{R}^n$  to  $\mathbb{R}$ ; they define a mapping  $F$  by

$$\mathbb{R}^n \ni x \mapsto F(x) := (f_1(x), \dots, f_m(x)) \in \mathbb{R}^m.$$

Equip  $\mathbb{R}^m$  with the dot-product and let  $g : \mathbb{R}^m \rightarrow \mathbb{R}$  be convex and increasing componentwise, i.e

$$y^i \geq z^i \text{ for } i = 1, \dots, m \implies g(y) \geq g(z).$$

Establishing the convexity of the function

$$\mathbb{R}^n \ni x \mapsto (g \circ F)(x) := g(f_1(x), \dots, f_m(x))$$

is an elementary exercise. Another easy observation is that, if  $(\rho^1, \dots, \rho^m) \in \partial g(y)$ , then each  $\rho^i$  is nonnegative: indeed,  $\{e_1, \dots, e_m\}$  being the canonical basis in  $\mathbb{R}^m$ ,

$$g(y) \geq g(y - e_j) \geq g(y) + \sum_{i=1}^m \rho^i (-e_j)^i = g(y) - \rho^j.$$

**Theorem 4.3.1** *Let  $f$ ,  $F$  and  $g$  be defined as above. For all  $x \in \mathbb{R}^n$ ,*

$$\begin{aligned} \partial(g \circ F)(x) = & \left\{ \sum_{i=1}^m \rho^i s_i : (\rho^1, \dots, \rho^m) \in \partial g(F(x)), \right. \\ & \left. s_i \in \partial f_i(x) \text{ for } i = 1, \dots, m \right\}. \end{aligned} \quad (4.3.1)$$

**PROOF.** [Preamble] Our aim is to show the formula via support functions, hence we need to establish the convexity and compactness of the right-hand side in (4.3.1) – call it  $S$ . Boundedness and closedness are easy, coming from the fact that a subdifferential (be it  $\partial g$  or  $\partial f_i$ ) is bounded and closed. As for convexity, pick two points in  $S$  and form their convex combination

$$s = \alpha \sum_{i=1}^m \rho^i s_i + (1 - \alpha) \sum_{i=1}^m \rho'^i s'_i = \sum_{i=1}^m \left[ \alpha \rho^i s_i + (1 - \alpha) \rho'^i s'_i \right],$$

where  $\alpha \in ]0, 1[$ . Remember that each  $\rho^i$  and  $\rho'^i$  is nonnegative and the above sum can be restricted to those terms such that  $\rho''^i := \alpha \rho^i + (1 - \alpha) \rho'^i > 0$ . Then we write each such term as

$$\rho''^i \left[ \frac{\alpha \rho^i}{\rho''^i} s_i + \frac{(1 - \alpha) \rho'^i}{\rho''^i} s'_i \right].$$

It suffices to observe that  $\rho''^i \in \partial g(F(x))$ , so the bracketed expression is in  $\partial f_i(x)$ ; thus  $s \in S$ .

[Step 1] Now let us compute the support function  $\sigma_S$  of  $S$ . For  $d \in \mathbb{R}^n$ , we denote by  $F'(x, d) \in \mathbb{R}^m$  the vector whose components are  $f'_i(x, d)$  and we proceed to prove

$$\sigma_S(d) = g'(F(x), F'(x, d)) . \quad (4.3.2)$$

For any  $s = \sum_{i=1}^m \rho^i s_i \in S$ , we write  $\langle s, d \rangle$  as

$$\sum_{i=1}^m \rho^i \langle s_i, d \rangle \leq \sum_{i=1}^m \rho^i f'_i(x, d) \leq g'(F(x), F'(x, d)) ; \quad (4.3.3)$$

the first inequality uses  $\rho^i \geq 0$  and the definition of  $f'_i(x, \cdot) = \sigma_{\partial f_i}(x)$ ; the second uses the definition  $g'(F(x), \cdot) = \sigma_{\partial g}(F(x))$ .

On the other hand, the compactness of  $\partial g(F(x))$  implies the existence of an  $m$ -tuple  $(\bar{\rho}^i) \in \partial g(F(x))$  such that

$$g'(F(x), F'(x, d)) = \sum_{i=1}^m \bar{\rho}^i f'_i(x, d) ,$$

and the compactness of each  $\partial f_i(x)$  yields likewise an  $\bar{s}_i \in \partial f_i(x)$  such that

$$f'_i(x, d) = \langle \bar{s}_i, d \rangle \quad \text{for } i = 1, \dots, m .$$

Altogether, we have exhibited an  $\bar{s} = \sum_{i=1}^m \bar{\rho}^i \bar{s}_i \in S$  such that equality holds in (4.3.3), so (4.3.2) is established.

[Step 2] It remains to prove that the support function (4.3.2) is really the directional derivative  $(g \circ F)'(x, d)$ . For  $t > 0$ , expand  $F(x + td)$ , use the fact that  $g$  is locally Lipschitzian, and then expand  $g(F(x + td))$ :

$$\begin{aligned} g(F(x + td)) &= g(F(x) + tF'(x, d) + o(t)) = g(F(x) + tF'(x, d)) + o(t) \\ &= g(F(x)) + tg'(F(x), F'(x, d)) + o(t) . \end{aligned}$$

From there, it follows

$$(g \circ F)'(x, d) := \lim_{t \downarrow 0} \frac{g(F(x + td)) - g(F(x))}{t} = g'(F(x), F'(x, d)) . \quad \square$$

Let us give some illustrations:

– When  $g$  is differentiable at  $F(x)$ , (4.3.1) has a classical flavour:

$$\partial(g \circ F)(x) = \sum_{i=1}^m \frac{\partial g}{\partial y^i}(F(x)) \partial f_i(x) .$$

In particular, with  $g(y^1, \dots, y^m) = \frac{1}{2} \sum_{i=1}^m (y^i)^+$  ( $r^+$  denoting  $\max\{0, r\}$ ), we obtain

$$\partial \left[ \frac{1}{2} \sum_{i=1}^m (f_i^+)^2 \right] = \sum_{i=1}^m f_i^+ \partial f_i .$$

– Take  $g(y^1, \dots, y^m) = \sum_{i=1}^m (y^i)^+$  and use the following notation:

$$I_0(x) := \{i : f_i(x) = 0\}, \quad I_+(x) := \{i : f_i(x) > 0\}.$$

Then

$$\partial \left[ \sum_{i=1}^m f_i^+ \right] (x) = \sum_{i \in I_+(x)} \partial f_i(x) + \sum_{i \in I_0(x)} [0, 1] \partial f_i(x).$$

– Finally, we give once more our fundamental example for optimization (generalizing Example 3.4, and to be generalized in §4.4):

**Corollary 4.3.2** *Let  $f_1, \dots, f_m$  be  $m$  convex functions from  $\mathbb{R}^n$  to  $\mathbb{R}$  and define*

$$f := \max \{f_1, \dots, f_m\}.$$

*Denoting by*

$$I(x) := \{i : f_i(x) = f(x)\}$$

*the active index-set, we have*

$$\partial f(x) = \text{co} \{\cup \partial f_i(x) : i \in I(x)\}. \quad (4.3.4)$$

PROOF. Take  $g(y) = \max\{y^1, \dots, y^m\}$ , whose subdifferential was computed in (3.7):  $\{e_i\}$  denoting the canonical basis of  $\mathbb{R}^m$ ,

$$\partial g(y) = \text{co}\{e_i : i \text{ such that } y^i = g(y)\}.$$

Then, using the notation of Theorem 4.3.1, we write  $\partial g(F(x))$  as

$$\left\{ (\rho^1, \dots, \rho^m) : \rho^i = 0 \text{ for } i \notin I(x), \quad \rho^i \geq 0 \text{ for } i \in I(x), \quad \sum_{i=1}^m \rho^i = 1 \right\},$$

and (4.3.1) gives

$$\partial f(x) = \left\{ \sum_{i \in I(x)} \rho^i \partial f_i(x) : \rho^i \geq 0 \text{ for } i \in I(x), \quad \sum_{i \in I(x)} \rho^i = 1 \right\}.$$

Remembering Example III.1.3.5, it suffices to recognize in the above expression the convex hull announced in (4.3.4)  $\square$

#### 4.4 Supremum of Convex Functions

We come now to an extremely important calculus rule, generalizing Corollary 4.3.2. It has no equivalent in classical differential calculus, and is of constant use in optimization. In this subsection, we study the following situation:  $J$  is an arbitrary index-set,  $\{f_j\}_{j \in J}$  is a collection of convex functions from  $\mathbb{R}^n$  to  $\mathbb{R}$ , and we assume that

$$f(x) := \sup \{f_j(x) : j \in J\} < +\infty \text{ for all } x \in \mathbb{R}^n. \quad (4.4.1)$$

We already know that  $f$  is convex (Proposition IV.2.1.2) and we are interested in computing its subdifferential. At a given  $x$ , call

$$J(x) := \{j \in J : f_j(x) = f(x)\} \quad (4.4.2)$$

the active index-set (possibly empty).

Let us start with an elementary result.

**Lemma 4.4.1** *With the notation (4.4.1), (4.4.2),*

$$\partial f(x) \supset \overline{\text{co}}\{\cup \partial f_j(x) : j \in J(x)\}. \quad (4.4.3)$$

PROOF. Take  $j \in J(x)$  and  $s \in \partial f_j(x)$ ; from the definition (1.2.1) of the subdifferential,

$$f(y) \geq f_j(y) \geq f_j(x) + \langle s, y - x \rangle \quad \text{for all } y \in \mathbb{R}^n,$$

so  $\partial f(x)$  contains  $\partial f_j(x)$ . Being closed and convex, it also contains the closed convex hull appearing in (4.4.3).  $\square$

Conversely, when is it true that the subdifferentials  $\partial f_j(x)$  at the active indices  $j$  “fill up” the whole of  $\partial f(x)$ ? This question is much more delicate, and requires some additional assumption, for example as follows:

**Theorem 4.4.2** *With the notation (4.4.1), (4.4.2), assume that  $J$  is a compact set (in some metric space), on which the functions  $j \mapsto f_j(x)$  are upper semi-continuous for each  $x \in \mathbb{R}^n$ . Then*

$$\partial f(x) = \text{co}\{\cup \partial f_j(x) : j \in J(x)\}. \quad (4.4.4)$$

PROOF. [Step 0] Our assumptions make  $J(x)$  nonempty and compact. Denote by  $S$  the curly bracketed set in (4.4.4); because of (4.4.3),  $S$  is bounded, let us check that it is closed. Take a sequence  $\{s_k\} \subset S$ , with  $\{s_k\}$  converging to  $s$ ; to each  $s_k$ , we associate some  $j_k \in J(x)$  such that  $s_k \in \partial f_{j_k}(x)$ , i.e.

$$f_{j_k}(y) \geq f_{j_k}(x) + \langle s_k, y - x \rangle \quad \text{for all } y \in \mathbb{R}^n.$$

Let  $k \rightarrow \infty$ ; extract a subsequence so that  $j_k \rightarrow j \in J(x)$ ; we have  $f_{j_k}(x) \equiv f(x) = f_j(x)$ ; and by upper semi-continuity of the function  $f(\cdot)(y)$ , we obtain

$$f_j(y) \geq \limsup f_{j_k}(y) \geq f_j(x) + \langle s, y - x \rangle \quad \text{for all } y \in \mathbb{R}^n,$$

which shows  $s \in \partial f_j(x) \subset S$ . Altogether,  $S$  is compact and its convex hull is also compact (Theorem III.1.4.3).

In view of Lemma 4.4.1, it suffices to prove the “ $\subset$ ”-inclusion in (4.4.4); for this, we will establish the corresponding inequality between support functions which, in view of the calculus rule V.3.3.3(ii), says: for all  $d \in \mathbb{R}^n$ ,

$$f'(x, d) \leq \sigma_S(d) = \sup \{f'_j(x, d) : j \in J(x)\}. \quad (4.4.5)$$

[Step 1] Let  $\varepsilon > 0$ ; from the definition (1.1.2) of  $f'(x, d)$ ,

$$\frac{f(x + td) - f(x)}{t} > f'(x, d) - \varepsilon \quad \text{for all } t > 0. \quad (4.4.6)$$

For  $t > 0$ , set

$$J_t := \left\{ j \in J : \frac{f_j(x + td) - f(x)}{t} \geq f'(x, d) - \varepsilon \right\}.$$

The definition of  $f(x + td)$  shows with (4.4.6) that  $J_t$  is nonempty. Because  $J$  is compact and  $f_{(\cdot)}(x + td)$  is upper semi-continuous,  $J_t$  is visibly compact. Observe that  $J_t$  is a superlevel-set of the function

$$0 < t \mapsto \frac{f_j(x + td) - f_j(x)}{t} + \frac{f_j(x) - f(x)}{t},$$

which is nondecreasing: the first fraction is the slope of a convex function, and the second fraction has a nonpositive numerator. Thus,  $J_{t_1} \subset J_{t_2}$  for  $0 < t_1 \leq t_2$ .

[Step 2] By compactness, we deduce the existence of some  $j^* \in \cap_{t>0} J_t$  (for each  $\tau \in ]0, t]$ , pick some  $j_\tau \in J_\tau \subset J_t$ ; take a cluster point for  $\tau \downarrow 0$ : it is in  $J_t$ ). We therefore have

$$f_{j^*}(x + td) - f(x) \geq t[f'(x, d) - \varepsilon] \quad \text{for all } t > 0,$$

hence  $j^* \in J(x)$  (continuity of the convex function  $f_{j^*}$  for  $t \downarrow 0$ ). In this inequality, we can replace  $f(x)$  by  $f_{j^*}(x)$ , divide by  $t$  and let  $t \downarrow 0$  to obtain

$$\sigma_S(d) \geq f'_{j^*}(x, d) \geq f'(x, d) - \varepsilon.$$

Since  $d \in \mathbb{R}^n$  and  $\varepsilon > 0$  were arbitrary, (4.4.5) is established.  $\square$

Some comments on the additional assumption are worth mentioning. First, the result concerns  $\partial f(x)$ , for which it is sufficient to know  $f$  only around  $x$ . It therefore applies if we have some neighborhood  $V$  of  $x$ , in which  $f$  is representable as

$$f(y) = \sup \{f_j(y) : j \in J(V)\} \quad \text{for all } y \in V,$$

where  $J(V)$  is a compact set on which  $j \mapsto f_j(y)$  is upper semi-continuous whenever  $y \in V$ . Secondly, this assumption deals with  $j$  only but this is somewhat misleading. The convexity of each  $f_j$  actually implies that  $f$  is *jointly* upper semi-continuous on  $J \times \mathbb{R}^n$ .

Finally, the set  $J$  is usually a subset of some  $\mathbb{R}^P$  and our assumption then implies three properties:  $J$  is closed, bounded, and the  $f_{(\cdot)}$  are upper semi-continuous. Let us examine what happens when *one* of these properties does not hold. If  $J$  is not closed, we may first have  $J(x) = \emptyset$ , in which case the formula is of no help. This case does not cause much trouble, though: nothing is changed if  $J$  is replaced by its closure, setting

$$\hat{f}_j(x) := \limsup_{j' \rightarrow j} f_{j'}(x)$$

for  $j \in (\text{cl } J) \setminus J$ . A simple example is

$$\mathbb{R} \ni x \mapsto f_j(x) = x - j \quad \text{with } j \in J = ]0, 1]. \quad (4.4.7)$$

Closing  $J$  places us in a situation in which applying Theorem 4.4.2 is trivial. The other two properties (upper semi-continuity and boundedness) are more fundamental.

**Example 4.4.3 [Upper Semi-Continuity]** Complete (4.4.7) by appending 0 to  $J$  and set  $f_0(x) \equiv 0$ ; then  $f(x) = x^+$  and

$$J(x) = \begin{cases} \{0\} & \text{if } x \leq 0, \\ \emptyset & \text{if } x > 0. \end{cases}$$

Here,  $j \mapsto f_j(x)$  is upper semi-continuous at  $x = 0$  only;  $J(0)$  yields  $\partial f_0(0) = \{0\} \subset \partial f(0) = [0, 1]$  and nothing more.

Observe that, introducing the upper semi-continuous hull  $\hat{f}_{(\cdot)}(x)$  of  $f_{(\cdot)}(x)$  does not change  $f$  (and hence  $\partial f$ ), but it changes the data, since the family now contains the additional function  $\hat{f}_0(x) = x^+$ . With the functions  $\hat{f}_j$  instead of  $f_j$ , formula (4.4.4) works.

[*Boundedness*] Take essentially the same functions as in the previous example, but with other notation:

$$f_0(x) \equiv 0, \quad f_j(x) = x - \frac{1}{j} \quad \text{for } j = 1, 2, \dots$$

Now  $J = \mathbb{N}$  is closed, and upper semi-continuity of  $f_{(\cdot)}(x)$  is automatic;  $f(x)$  and  $J(x)$  are as before and the same discrepancy occurs at  $x = 0$ .  $\square$

A special case of Theorem 4.4.2 is when each  $f_j$  is differentiable (see Corollary 2.1.4 to remember what it means exactly).

**Corollary 4.4.4** *The notation and assumptions are those of Theorem 4.4.2. Assume also that each  $f_j$  is differentiable; then*

$$\partial f(x) = \text{co} \{\nabla f_j(x) : j \in J(x)\}. \quad \square$$

A geometric proof of this result was given in Example 3.4, in the simpler situation of finitely many affine functions  $f_j$ . Thus, in the framework of Corollary 4.4.4, and whenever there are only finitely many active indices at  $x$ ,  $\partial f(x)$  is a compact convex polyhedron, generated by the active gradients at  $x$ .

The case of  $J(x)$  being a singleton deserves further comments. We rewrite Corollary 4.4.4 in this case, using a different notation reflecting a situation frequently encountered in optimization.

**Corollary 4.4.5** *For some compact set  $Y \subset \mathbb{R}^p$ , let  $g : \mathbb{R}^n \times Y \rightarrow \mathbb{R}$  be a function satisfying the following properties:*

- for each  $x \in \mathbb{R}^n$ ,  $g(x, \cdot)$  is upper semi-continuous;
- for each  $y \in Y$ ,  $g(\cdot, y)$  is convex and differentiable;
- the function  $f := \sup_{y \in Y} g(\cdot, y)$  is finite-valued on  $\mathbb{R}^n$ ;
- at some  $x \in \mathbb{R}^n$ ,  $g(x, \cdot)$  is maximized at a unique  $y(x) \in Y$ .

*Then  $f$  is differentiable at this  $x$ , and its gradient is*

$$\nabla f(x) = \nabla_x g(x, y(x)) \tag{4.4.8}$$

*(where  $\nabla_x g(x, y)$  denotes the gradient of the function  $g(\cdot, y)$  at  $x$ ).*  $\square$

Computing  $f$  at a given  $x$  amounts to solving a certain maximization problem, to obtain a solution  $y^*$ , say (which depends on  $x$ !). Then a practical rule is: to obtain the gradient of  $f$ , simply differentiate  $g$  with respect to  $x$ , the variable  $y$  being set to this value  $y^*$ . If, by any chance, no other  $y \in Y$  maximizes  $g$  at this  $x$ , one does get  $\nabla f(x)$ . If not, at least a subgradient is obtained (Lemma 4.4.1).

**Remark 4.4.6** When  $Y$  is a finite set, Corollary 4.4.5 can be easily accepted (see Corollary 4.3.2): when  $x$  varies,  $y^*$  stays locally the same, just because each  $g(\cdot, y)$  is continuous.

When  $Y$  is infinite, however, a really baffling phenomenon occurs: although  $f$  is a fairly complicated function, its gradient exists (!), and is given by the very simple formula (4.4.8) (!!). It is perhaps easier to accept this result after looking at the following naive calculation.

Suppose that we are in a very favourable situation:  $Y$  is some space in which differentiation is possible;  $g(\cdot, \cdot)$  is a smooth function; the problem  $\max g(x + h, \cdot)$  has a unique solution for  $h$  close to 0; and finally, this unique solution  $y(\cdot)$  is a smooth function. Then write formally

$$\nabla f(x) = \nabla_x g(x, y(x)) + \nabla_y g(x, y(x)) \nabla y(x)$$

and here comes the trickery: because  $y(x)$  is a maximal point,  $g(x, \cdot)$  is stationary at  $y(x)$ ; the second term is therefore zero.  $\square$

Because of the importance of sup-functions, we give one more result, valid *without any assumption*, in which case (4.4.4) breaks down from the very beginning. When the crucial property  $J(x) \neq \emptyset$  does not hold, a natural cure is to enlarge  $J(x)$ , so as to take into account the indices that are almost active; we therefore set, for given  $x$  and  $\delta > 0$ ,

$$J_\delta(x) := \{j \in J : f_j(x) \geq f(x) - \delta\}. \quad (4.4.9)$$

This is not enough, however, as shown by the following counter-example with  $n = 1$ :

$$f_j(x) = |x|^j \quad \text{for } j \in J = [1, 2]. \quad (4.4.10)$$

Then  $f(x) = |x|$  for  $x \in [-1, +1]$  (and  $x^2$  elsewhere). At  $x = 0$ , every  $j \in J$  is active, and considering the almost active indices brings just nothing; but every  $\partial f_j(0)$  is reduced to  $\{0\}$ ! Some further idea is wanted; this idea is to collect also the subgradients around the given point – which we will now call  $x_0$ . Thus, we are interested in the set

$$S_\delta := \cup\{\partial f_j(x) : j \in J_\delta(x_0), x \in B(x_0, \delta)\}. \quad (4.4.11)$$

To recover  $\partial f(x)$ , we will simply let  $\delta \downarrow 0$ ; everything is now set to obtain the most possible general formula in the present framework of a finite-valued sup-function. First, we need a technical lemma.

**Lemma 4.4.7** *For given  $x_0 \in \mathbb{R}^n$  and  $\delta > 0$ , consider the following index-set:*

$$J^* := \cup\{J_\delta(x) : x \in B(x_0, 2\delta)\}.$$

*There exists a common Lipschitz constant  $L$  for the functions  $\{f_j\}_{j \in J^*}$  on the ball  $B(x_0, \delta)$ , and for  $f$  on  $B(x_0, 2\delta)$ .*

*As a result, the set  $S_\delta$  of (4.4.11) is bounded and, for  $s \in S_\delta$ , there holds*

$$f(y) \geq f(x_0) + \langle s, y - x_0 \rangle - (4L + 2)\delta \quad \text{for all } y \in \mathbb{R}^n. \quad (4.4.12)$$

PROOF. Because  $f$  is finite-valued, there are  $m$  and  $M$  such that

$$m \leq f(x) \leq M \quad \text{for all } x \in B(x_0, 4\delta)$$

and therefore

$$m - \delta \leq f_j(x) \leq M \quad \text{for all } (j, x) \in J^* \times B(x_0, 2\delta).$$

Then the Lipschitz properties stated follow from Lemma IV.3.1.1.

In particular, take  $s \in S_\delta$ : there are  $j \in J_\delta(x_0) \subset J^*$  and  $x \in B(x_0, \delta)$  such that  $s \in \partial f_j(x)$ ; write the following chain of inequalities:

$$\begin{aligned} \langle s, y - x \rangle &\leq f_j(y) - f_j(x) \leq f(y) - f_j(x) \leq \\ &\leq f(y) - f_j(x_0) + L\delta \leq f(y) - f(x_0) + (L+1)\delta. \end{aligned} \quad (4.4.13)$$

This is true for all  $y$ ; if  $s \neq 0$ , take  $y = x + \delta s / \|s\| \in B(x_0, 2\delta)$  and use the Lipschitz property of  $f$  to obtain

$$\delta \|s\| \leq 2L\delta + (L+1)\delta.$$

A bound is thus established for  $s$ . Use it in (4.4.13):

$$\langle s, y - x_0 \rangle \leq f(y) - f(x_0) + (L+1)\delta + \|s\|\delta \leq f(y) - f(x_0) + (4L+2)\delta. \quad \square$$

**Theorem 4.4.8** *With the notation (4.4.1), (4.4.9), (4.4.11) and given  $x_0 \in \mathbb{R}^n$ ,*

$$\partial f(x_0) = \bigcap_{\delta > 0} \overline{\text{co}} S_\delta. \quad (4.4.14)$$

PROOF.  $[ \supset ]$  If  $s \in S_\delta$ , we know that (4.4.12) holds; it holds also for all convex combinations, and for all limits of such. If  $s \in \overline{\text{co}} S_\delta$  for all  $\delta > 0$ ,  $s$  therefore satisfies (4.4.12) for all  $\delta > 0$ , and is thus in  $\partial f(x_0)$ .

$[ \subset ]$  The right-hand side in (4.4.14) is nonempty, being an intersection of nested nonempty compact sets. We use support functions: from the calculus rule V.3.3.3(iii), we need to show that  $f'(x, \cdot) \leq \inf_{\delta > 0} \sigma_{S_\delta}$ . Choose  $\delta > 0$ ,  $\varepsilon > 0$ , and  $d$  of norm 1. Since

$$\sigma_{S_\delta}(d) = \sup \{f'_j(x, d) : j \in J_\delta(x_0), x \in B(x_0, \delta)\},$$

we will be done if we single out  $j^* \in J_\delta(x_0)$  and  $x^* \in B(x_0, \delta)$  such that

$$f'(x_0, d) \leq f'_{j^*}(x^*, d) + \varepsilon.$$

With the notations of Lemma 4.4.7, we choose first  $0 < t^* \leq \min\{\varepsilon, \delta\}$  such that  $2Lt^* + t^{*2} \leq \delta$ . Then we choose  $j^* \in J$  such that

$$f(x_0 + t^*d) \leq f_{j^*}(x_0 + t^*d) + t^{*2}.$$

Because  $t^{*2} \leq \delta$  and  $x_0 + t^*d \in B(x_0, \delta)$ ,  $j^* \in J^*$  and the Lipschitz properties allow us to write

$$f(x_0) - Lt^* \leq f_{j^*}(x_0 + t^*d) + t^{*2} \leq f_{j^*}(x_0) + Lt^* + t^{*2};$$

we do have  $j^* \in J_\delta(x_0)$ .

On the other hand,

$$\begin{aligned} f'(x_0, d) &\leq \frac{f(x_0 + t^*d) - f(x_0)}{t^*} \\ &\leq \frac{f_{j^*}(x_0 + t^*d) + t^{*2} - f(x_0)}{t^*} \\ &\leq \frac{f_{j^*}(x_0 + t^*d) - f_{j^*}(x_0)}{t^*} + t^* \leq f'_{j^*}(x^*, d) + \varepsilon, \end{aligned}$$

where we have used the mean-value Theorem 2.3.3:  $x^* \in ]x_0, x_0 + t^*d[ \subset B(x_0, \delta)$ . In summary, our  $j^*$  and  $x^*$  do satisfy the required properties, the theorem is proved.  $\square$

Apart from its ability to describe  $\partial f(x_0)$  accurately, this result gives a practical alternative to Lemma 4.4.1: given  $x_0$ , first compute some  $j$  solving approximately the optimization problem (4.4.1); then compute a subgradient of  $f_j$ , possibly at some neighboring point  $x$ ; this subgradient is reasonably close to  $\partial f(x_0)$  – remember in particular (4.4.12).

#### 4.5 Image of a Function Under a Linear Mapping

Let  $g : \mathbb{R}^m \rightarrow \mathbb{R}$  be a convex function and  $A : \mathbb{R}^m \rightarrow \mathbb{R}^n$  a surjective linear operator. We recall from §IV.2.4 that the associated function

$$\mathbb{R}^n \ni x \mapsto (Ag)(x) := \inf \{g(y) : Ay = x\} \quad (4.5.1)$$

is convex, provided that, for all  $x$ ,  $g$  is bounded from below on  $A^{-1}(x)$ . Analogously to (4.4.2), we denote by

$$Y(x) := \{y \in \mathbb{R}^m : Ay = x, g(y) = (Ag)(x)\} \quad (4.5.2)$$

the set of minimizers in (4.5.1).

**Theorem 4.5.1** *With the notation (4.5.1), (4.5.2), let  $x$  be such that  $Y(x)$  is nonempty. Then, for arbitrary  $y \in Y(x)$ ,*

$$\partial(Ag)(x) = \{s \in \mathbb{R}^n : A^*s \in \partial g(y)\} = (A^*)[\partial g(y)] \quad (4.5.3)$$

(and this set is thus independent of the particular optimal  $y$ ).

PROOF. By definition,  $s \in \partial(Ag)(x)$  if and only if

$$(Ag)(x') \geq (Ag)(x) + \langle s, x' - x \rangle \quad \text{for all } x' \in \mathbb{R}^n,$$

which can be rewritten

$$(Ag)(x') \geq g(y) + \langle s, x' - Ay \rangle \quad \text{for all } x' \in \mathbb{R}^n$$

where  $y$  is arbitrary in  $Y(x)$ . Furthermore, because  $A$  is surjective and by definition of  $Ag$ , this last relation is equivalent to

$$g(y') \geq g(y) + \langle s, Ay' - Ay \rangle = g(y) + \langle A^*s, y' - y \rangle \quad \text{for all } y' \in \mathbb{R}^m$$

which means that  $A^*s \in \partial g(y)$ .  $\square$

The surjectivity of  $A$  implies first that  $(Ag)(x) < +\infty$  for all  $x$ , but it has a more interesting consequence:

**Corollary 4.5.2** *In (4.5.1), (4.5.2), if  $g$  is differentiable at some  $y \in Y(x)$ , then  $Ag$  is differentiable at  $x$ .*

PROOF. Surjectivity of  $A$  is equivalent to injectivity of  $A^*$ : in (4.5.3), we have an equation in  $s$ :  $A^*s = \nabla g(y)$ , whose solution is unique, and is therefore  $\nabla(Ag)(x)$ .  $\square$

A first example of image-function is when  $A$  is a restriction in a product space:  $g$  being a convex function on  $\mathbb{R}^n \times \mathbb{R}^m$ , consider the marginal function, obtained by partial minimization of  $g$ :

$$\mathbb{R}^n \ni x \mapsto f(x) := \inf \{g(x, y) : y \in \mathbb{R}^m\}. \quad (4.5.4)$$

This  $f$  is put under the form  $Ag$ , if we choose  $A : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$  defined by  $A(x, y) = x$ .

**Corollary 4.5.3** *Suppose that the subdifferential of  $g$  in (4.5.4) is associated with a scalar product  $\langle \cdot, \cdot \rangle$  preserving the structure of a product space: for all  $x, x' \in \mathbb{R}^n$  and  $y, y' \in \mathbb{R}^m$ ,*

$$\langle (x, y), (x', y') \rangle = \langle x, x' \rangle_n + \langle y, y' \rangle_m.$$

*At a given  $x \in \mathbb{R}^n$ , take an arbitrary  $y$  solving (4.5.4). Then*

$$\partial f(x) = \{s \in \mathbb{R}^n : (s, 0) \in \partial_{(x,y)}g(x, y)\}.$$

PROOF. With our notation,  $A^*s = (s, 0)$  for all  $s \in \mathbb{R}^n$ . It suffices to apply Theorem 4.5.1 (the symbol  $\partial_{(x,y)}g$  is used as a reminder that we are dealing with the subdifferential of  $g$  with respect to the variable  $(\cdot, \cdot) \in \mathbb{R}^n \times \mathbb{R}^m$ ).  $\square$

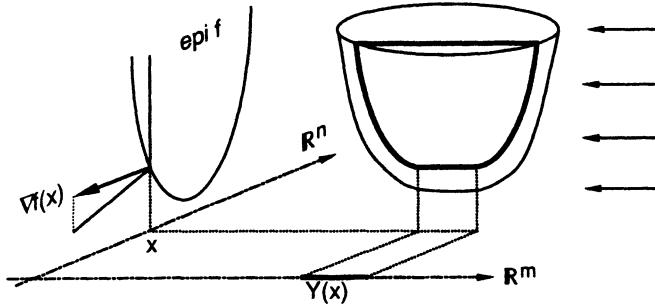
If  $g$  is differentiable on  $\mathbb{R}^n \times \mathbb{R}^m$  and is minimized “at finite distance” in (4.5.4), then the resulting  $f$  is differentiable (see Remark 4.5.2). In fact,

$$\nabla_{(x,y)}g(x, y) = (\nabla_x g(x, y), \nabla_y g(x, y)) \in \mathbb{R}^n \times \mathbb{R}^m$$

and the second component is 0 just because  $y$  is a minimizer. We do obtain

$$\nabla f(x) = \nabla_x g(x, y) \quad \text{with } y \text{ solving (4.5.4).}$$

A geometric explanation of this differentiability property appears on Fig. 4.5.1: the shadow of a smooth convex epigraph is normally a smooth convex epigraph.



**Fig. 4.5.1.** The gradient of a marginal function

**Remark 4.5.4** The following counter-example emphasizes the necessity of a minimizer  $y$  to apply Theorem 4.5.1: in  $\mathbb{R}^2$ , the function

$$g(x, y) := \sqrt{x^2 + e^{2y}}$$

is convex (check it), perfectly smooth ( $C^\infty$ ), but “minimal at infinity” (for all  $x$ ). The resulting marginal function  $f(x) = |x|$  is not a smooth function.  $\square$

Another important instance of an image-function was seen in §IV.2.3: the infimal convolution of two functions, defined by

$$(f_1 \downarrow f_2)(x) := \inf \{f_1(y_1) + f_2(y_2) : y_1, y_2 \in \mathbb{R}^n, y_1 + y_2 = x\}. \quad (4.5.5)$$

Recall from the end of §IV.2.4 that this operation can be put in the form (4.5.1), by considering

$$\begin{aligned} \mathbb{R}^n \times \mathbb{R}^n &\ni (y_1, y_2) \mapsto g(y_1, y_2) := f_1(y_1) + f_2(y_2) \in \mathbb{R}, \\ \mathbb{R}^n \times \mathbb{R}^n &\ni (y_1, y_2) \mapsto A(y_1, y_2) := y_1 + y_2 \in \mathbb{R}^n. \end{aligned}$$

**Corollary 4.5.5** Let  $f_1$  and  $f_2 : \mathbb{R}^n \rightarrow \mathbb{R}$  be two convex functions minorized by a common affine function. For given  $x$ , let  $(y_1, y_2)$  be such that the inf-convolution is exact at  $x = y_1 + y_2$ , i.e.:  $(f_1 \downarrow f_2)(x) = f_1(y_1) + f_2(y_2)$ . Then

$$\partial(f_1 \downarrow f_2)(x) = \partial f_1(y_1) \cap \partial f_2(y_2). \quad (4.5.6)$$

**PROOF.** First observe that  $A^*s = (s, s)$ . Also, apply Definition 1.2.1 to see that  $(s_1, s_2) \in \partial g(y_1, y_2)$  if and only if  $s_1 \in \partial f_1(y_1)$  and  $s_2 \in \partial f_2(y_2)$ . Then (4.5.6) is just the copy of (4.5.3) in the present context.  $\square$

Once again, we obtain a regularity result (among others):  $\nabla(f_1 \downarrow f_2)(x)$  exists whenever there is an optimal  $(y_1, y_2)$  in (4.5.5) for which either  $f_1$  or  $f_2$  is differentiable. For an illustration, see again Example IV.2.3.8, more precisely (IV.2.3.6).

**Remark 4.5.6** In conclusion, let us give a warning: the max-operation (§4.4) does not destroy differentiability if uniqueness of the argmax holds. By contrast, the differentiability of a min-function (§4.5) has nothing to do with uniqueness of the argmin.

This may seem paradoxical, since maximization and minimization are just the same operation, as far as *differentiability* of the result  $f$  is concerned. Observe, however, that the differentiability obtained in §4.5 relies heavily on the *joint* convexity of the underlying function  $g$  of Theorem 4.5.1 and Corollary 4.5.3. This last property has little relevance in §4.4.  $\square$

## 5 Further Examples

With the help of the calculus rules developed in §4, we can study more sophisticated examples than those of §3. They will reveal, in particular, the important role of §4.4: max-functions appear all the time.

### 5.1 Largest Eigenvalue of a Symmetric Matrix

We adopt the notation of §IV1.3(e): in the space  $S_n(\mathbb{R})$  of symmetric  $n \times n$  matrices equipped with  $\langle\!\langle \cdot, \cdot \rangle\!\rangle$ , the function

$$S_n(\mathbb{R}) \ni M \mapsto \lambda_1(M) \in \mathbb{R}$$

is convex and can be represented as

$$\lambda_1(M) = \max \{ u^\top M u : u \in \mathbb{R}^n, u^\top u = 1 \}. \quad (5.1.1)$$

Furthermore, the set of optimal  $u$  in (5.1.1) is the set of normalized eigenvectors associated with the resulting  $\lambda_1$ .

Thus, Corollary 4.4.4 will give the subdifferential of  $\lambda_1$  if the gradient of the function  $M \mapsto u^\top M u$  can be computed. This is easy: by direct calculation, we have

$$u^\top M u = \langle\!\langle uu^\top, M \rangle\!\rangle \quad (5.1.2)$$

so the linear function  $u \mapsto u^\top M u$  supports the singleton  $\{u^\top u\}$ , a rank-one matrix of  $S_n(\mathbb{R})$  (its kernel is the subspace orthogonal to  $u$ ). The subdifferential of  $\lambda_1$  at  $M$  is therefore the convex hull of all these matrices:

$$\partial \lambda_1(M) = \text{co} \{ uu^\top : u^\top u = 1, Mu = \lambda_1(M)u \}. \quad (5.1.3)$$

Naturally, this is the face of  $\partial \lambda_1(0)$  exposed by  $M$ , where  $\partial \lambda_1(0)$  was given in Example V.3.3.11. It is the singleton  $\{\nabla \lambda_1(M)\}$  if and only if the maximal eigenvalue  $\lambda_1$  of  $M$  is simple.

The directional derivatives of  $\lambda_1$  can of course be computed: the support function of (5.1.3) is, using (5.1.2) to reduce superfluous notation,

$$P \mapsto \lambda'_1(M, P) = \max \{ u^\top P u : u \text{ normalized eigenvector for } \lambda_1(M) \}.$$

**Remark 5.1.1** It is tempting to think of the problem as follows. There are a finite number of eigenvalues; each one is a root of the characteristic polynomial of  $M$ , whose coefficients are

smooth functions of  $M$ ; therefore, each eigenvalue is a smooth function of  $M$ :  $\lambda_1$  is a max of finitely many smooth functions.

If this reasoning were correct, Corollary 4.4.4 would tell us that  $\partial\lambda_1(M)$  is a compact convex polyhedron; this is certainly not the case of the set (5.1.3)! The flaw is that the roots of a polynomial cannot be enumerated. When they are all distinct, each can be followed by continuity; but when two roots coincide, this continuity argument vanishes.  $\square$

As an example, let us study the cone of symmetric negative semi-definite matrices, i.e. the sublevel-set

$$K^- := \{M : \lambda_1(M) \leq 0\}.$$

Its boundary is the set of matrices  $M \in K^-$  that are singular ( $-I_n$  has  $\lambda_1(-I_n) < 0$ , so Proposition 1.3.3 applies). For such  $M$ , Proposition 1.3.4 characterizes the tangent and normal cones to  $K^-$  at  $M$ :

$$\begin{aligned} T_{K^-}(M) &= \{P : u^\top P u \leq 0 \text{ for all } u \in \text{Ker } M\}, \\ N_{K^-}(M) &= \text{co}\{uu^\top : u \in \text{Ker } M\}. \end{aligned}$$

If  $M \neq 0$ , Example III.5.2.6 gives a more handy expression for the normal cone:

$$N_{K^-}(M) = \{P \text{ symmetric positive semi-definite} : \langle\langle M, P \rangle\rangle = 0\}.$$

In problems involving largest eigenvalues, the variable matrix  $M$  is often imposed a certain pattern. For example, one considers matrices with fixed off-diagonal elements, only their diagonal being free. In that case, a symmetric matrix  $M_0$  is given and the function to be studied is  $\lambda_1(M_0 + D)$ , where  $D$  is an arbitrary diagonal  $n \times n$  matrix. Identifying the set of such diagonals with  $\mathbb{R}^n$ , we thus obtain the function

$$f(x) = f(\xi^1, \dots, \xi^n) := \lambda_1(M_0 + \text{diag}(\xi^1, \dots, \xi^n)).$$

This  $f$  is  $\lambda_1$  pre-composed with an affine mapping whose linear part is  $A_0 : \mathbb{R}^n \rightarrow S_n(\mathbb{R})$  defined by

$$\mathbb{R}^n \ni x = (\xi^1, \dots, \xi^n) \mapsto A_0(x) := \text{diag}(\xi^1, \dots, \xi^n) \in S_n(\mathbb{R}).$$

We have

$$\langle\langle A_0 x, M \rangle\rangle = \sum_{i=1}^n \xi^i M_{ii} \quad \text{for all } x \in \mathbb{R}^n \text{ and } M \in S_n(\mathbb{R}).$$

Knowing that  $\mathbb{R}^n$  is equipped with the usual dot-product, the adjoint of  $A_0$  is therefore defined by

$$x^\top A_0^* M = \sum_{i=1}^n \xi^i M_{ii} \quad \text{for all } x \in \mathbb{R}^n \text{ and } M \in S_n(\mathbb{R}).$$

Thus,  $A_0^* : S_n(\mathbb{R}) \rightarrow \mathbb{R}^n$  appears as the operator that takes an  $n \times n$  matrix and makes an  $n$ -vector with its diagonal elements. Because the  $(i, j)^{\text{th}}$  element of the matrix  $uu^\top$  is  $u^i u^j$ , (5.1.3) gives with the calculus rule (4.2.1)

$$\partial f(x) = \text{co}\left\{((u^1)^2, \dots, (u^n)^2) : u \text{ normalized eigenvector at } f(x)\right\}.$$

## 5.2 Nested Optimization

In optimization, convex functions that are themselves the result of some other optimization problem are encountered fairly often. Let us mention, among others, problems issuing from game theory, all kinds of decomposition schemes, semi-infinite programming, optimal control problems in which the state equation is replaced by an inclusion, etc. We consider two examples below, which are still within the framework of this book: partially linear least-squares problems, and Lagrangian relaxation.

**(a) Partially Linear Least-Squares Problems** In our first example, there are three vector spaces:  $\mathbb{R}^n$ ,  $\mathbb{R}^m$  and  $\mathbb{R}^P$ , each equipped with its dot-product. A matrix  $A(x) : \mathbb{R}^m \rightarrow \mathbb{R}^P$  is given, depending on the parameter  $x \in \mathbb{R}^n$ , as well as a vector  $b(x) \in \mathbb{R}^P$ . Then one considers the function

$$\mathbb{R}^n \times \mathbb{R}^m \ni (x, y) \mapsto g(x, y) := \frac{1}{2} \|A(x)y - b(x)\|^2. \quad (5.2.1)$$

The problem is to minimize  $g$ , assumed for simplicity to be convex.

It makes good sense to minimize  $g$  hierarchically: first with respect to  $y$  ( $x$  being fixed), and then minimize the result with respect to  $x$ . In other words, defining

$$f(x) := \min \{g(x, y) : y \in \mathbb{R}^m\},$$

(5.2.1) is replaced by the problem of minimizing  $f$  with respect to  $x$ . In theory, nothing is changed; in practice, a lot is changed.

For one thing,  $f$  has less variables than  $g$ . More importantly, however, it is usually the case in the model (5.2.1) that  $x$  and  $y$  have nothing to do with each other. For example,  $y$  may be a set of weights, measured in kilograms; and  $x$  may be interest rates, i.e. dimensionless numbers. Under these conditions, any numerical method to minimize  $g$  directly will run into trouble because an appropriate *scaling* is hard to find. To cut a long story short,  $f$  is likely to be more easily minimized than  $g$ .

Now,  $y$  is given by a linear least-squares system

$$A^\top(x)[A(x)y - b(x)] = 0 \quad (5.2.2)$$

which has always a solution (not necessarily unique): we are right in the framework of Corollary 4.5.3. Without any assumption on the rank of  $A(x)$ ,

$$\nabla f(x) = [A'(y) - b'](Ay - b). \quad (5.2.3)$$

Here  $y$  is any solution of (5.2.2);  $b'$  is the matrix whose  $k^{\text{th}}$  row is the derivative of  $b$  with respect to the  $k^{\text{th}}$  component  $\xi^k$  of  $x$ ;  $A'(y)$  is the matrix whose  $k^{\text{th}}$  row is  $y^\top(A'_k)^\top$ ;  $A'_k$  is the derivative of  $A$  with respect to  $\xi^k$ . Then,  $f$  can be minimized numerically by any of the available algorithms of Chap. II, having (5.2.3) as the black box (U1) of Fig. II.1.2.1. It is most probable that a very efficient method will thus be obtained.

**(b) Lagrangian Relaxation** Our second example is of utmost practical importance and will motivate the full Chapter XII. Here, we content ourselves with a brief description of the problem. Given a set  $U$  and  $n + 1$  functions  $c_0, c_1, \dots, c_n$  from  $U$  to  $\mathbb{R}$ , consider the problem

$$\begin{cases} \sup c_0(u) & u \in U, \\ c_i(u) = 0 & \text{for } i = 1, \dots, n. \end{cases} \quad (5.2.4)$$

Associated with this problem is the *Lagrange function*, which depends on  $x = (\xi^1, \dots, \xi^n) \in \mathbb{R}^n$  and  $u \in U$ , and is defined by

$$g(x, u) := c_0(u) + \sum_{i=1}^n \xi^i c_i(u).$$

We will see in Chap. XII that another function is important for solving (5.2.4), namely

$$f(x) := \sup \{g(x, u) : u \in U\}, \quad (5.2.5)$$

which must be minimized. Needless to say,  $f$  is convex, as the supremum of the linear functions  $g(\cdot, u)$ . In the good cases, when the hypotheses of Theorem 4.4.2 bring no trouble, its subdifferential is given by Corollary 4.4.4

$$\partial f(x) = \text{co} \{c(u) : u \in U(x)\} \quad (5.2.6)$$

where  $c(u) \in \mathbb{R}^n$  denotes the vector whose coordinates are  $c_i(u)$ , and  $U(x)$  is the optimal set in (5.2.5).

According to (5.2.6), the subgradients of  $f$  are obtained from the constraint-values at those  $u$  solving (5.2.5); at least, the inclusion “ $\supset$ ” always holds in (5.2.6), and approximations are possible if Theorem 4.4.8 must be invoked. An  $\bar{x}$  minimizing  $f$  is characterized by the following condition: for some positive integer  $p \leq n + 1$ , there exist  $u_1, \dots, u_p$  in  $U(\bar{x})$  and a set of convex multipliers  $\alpha = (\alpha_1, \dots, \alpha_p) \in \Delta_p$  such that

$$g(\bar{x}, u_k) = f(\bar{x}) \quad \text{and} \quad \sum_{k=1}^p \alpha_k c(u_k) = 0 \in \mathbb{R}^n.$$

In particular, if  $g(\bar{x}, \cdot)$  happens to have a unique maximum  $\bar{u}$ , then  $p = 1$ , which means that  $c(\bar{u}) = 0$ .

At a non-optimal  $x$ , the descent directions for  $f$  are described by Theorem 1.3.4:

$$\text{int T}_{Sf(x)}(x) = \{d \in \mathbb{R}^n : d^\top c(u) < 0 \text{ for all } u \in U(x)\}.$$

### 5.3 Best Approximation of a Continuous Function on a Compact Interval

Let  $T$  be a compact interval of  $\mathbb{R}$  and  $\varphi_0$  a real-valued continuous function defined on  $T$ . Furthermore,  $n$  functions  $\varphi_1, \dots, \varphi_n$  are given in the space  $C(T)$  of real-valued continuous functions on  $T$ ; usually, they are linearly independent. We are interested in finding a linear combination of the  $\varphi_i$ 's which best approximates  $\varphi_0$ , in the sense of the max-norm. In other words, we want to minimize over  $\mathbb{R}^n$  the *error-function*

$$f(x) := \max \{|g(x, t)| : t \in T\} \quad (5.3.1)$$

where  $g$  denotes the function (affine in  $x$ )

$$g(x, t) := \sum_{i=1}^n \xi^i \varphi_i(t) - \varphi_0(t) = [\varphi(t)]^\top x - \varphi_0(t). \quad (5.3.2)$$

Minimizing  $f$  is one of the simplest instances of *semi-infinite programming*: optimization problems with finitely many variables but infinitely many constraints.

The error-function is convex and, once more, enters the framework of Corollary 4.4.4 (observe in particular that  $|g| = \max\{g, -g\}$ ). We neglect the case of an  $x$  with  $f(x) = 0$ , which is not particularly interesting:  $g(x, \cdot) \equiv 0$ ,  $x$  is optimal anyway. Denoting by  $H$  the (usually  $n$ -dimensional) subspace of  $C(T)$  generated by the  $\varphi_i$ 's, we therefore assume  $\varphi_0 \notin H$ .

Fix an  $x$  with  $f(x) > 0$ , and call  $T(x) \subset T$  the set of  $t$  yielding the max in (5.3.1);  $T(x)$  is nonempty from our assumptions. At each such  $t$ , we can define  $\varepsilon(t) \in \{-1, +1\}$  by

$$\varepsilon(t)g(x, t) = f(x) \quad \text{for all } t \in T(x).$$

Then,  $\partial f(x)$  is the convex combination of all the  $n$ -vectors  $\varepsilon(t)\varphi(t)$ , where  $t$  describes  $T(x)$ . Deriving an optimality condition is then easy with Corollary 4.4.4 and Theorem 4.1.1:

**Theorem 5.3.1** *With the notations (5.3.1), (5.3.2), suppose  $\varphi_0 \notin H$ . A necessary and sufficient condition for  $\bar{x} = (\bar{\xi}^1, \dots, \bar{\xi}^n) \in \mathbb{R}^n$  to minimize  $f$  of (5.3.1) is that, for some positive integer  $p \leq n+1$ , there exist  $p$  points  $t_1, \dots, t_p$  in  $T$ ,  $p$  integers  $\varepsilon_1, \dots, \varepsilon_p$  in  $\{-1, +1\}$  and  $p$  positive numbers  $\alpha_1, \dots, \alpha_p$  such that*

$$\begin{aligned} \sum_{i=1}^n \bar{\xi}^i \varphi_i(t_k) - \varphi_0(t_k) &= \varepsilon_k f(\bar{x}) \quad \text{for } k = 1, \dots, p, \\ \sum_{k=1}^p \alpha_k \varepsilon_k \varphi_i(t_k) &= 0 \quad \text{for } i = 1, \dots, n \\ (\text{or equivalently: } \sum_{k=1}^p \alpha_k \varepsilon_k \psi(t_k) &= 0 \quad \text{for all } \psi \in H). \end{aligned} \quad \square$$

Indeed, this example is formally identical to Lagrangian relaxation; the possible differences are usually in the assumptions on  $T$ , which plays the role of  $U$ .

## 6 The Subdifferential as a Multifunction

Section 2 was mainly concerned with properties of the “static” set  $\partial f(x)$ . Here, we study the properties of this set varying with  $x$ , and also with  $f$ .

## 6.1 Monotonicity Properties of the Subdifferential

We have seen in §IV.4.1 that the gradient mapping of a differentiable convex function is *monotone*, a concept generalizing to several dimensions that of a nondecreasing function. Now, this monotonicity has its formulation even in the absence of differentiability.

**Proposition 6.1.1** *The subdifferential mapping is monotone in the sense that, for all  $x_1$  and  $x_2$  in  $\mathbb{R}^n$ ,*

$$\langle s_2 - s_1, x_2 - x_1 \rangle \geq 0 \quad \text{for all } s_i \in \partial f(x_i), i = 1, 2. \quad (6.1.1)$$

PROOF. The subgradient inequalities

$$\begin{aligned} f(x_2) &\geq f(x_1) + \langle s_1, x_2 - x_1 \rangle \quad \text{for all } s_1 \in \partial f(x_1) \\ f(x_1) &\geq f(x_2) + \langle s_2, x_1 - x_2 \rangle \quad \text{for all } s_2 \in \partial f(x_2) \end{aligned}$$

give the result simply by addition.  $\square$

A convex function can be “more or less non-affine”, according to how much its graph deviates from a hyperplane. We recall, for example, that  $f$  is strongly convex on a convex set  $C$  when, for some modulus of strong convexity  $c > 0$ , all  $x_1, x_2$  in  $C$ , and all  $\alpha \in ]0, 1[$ , it holds

$$f(\alpha x_2 + (1 - \alpha)x_1) \leq \alpha f(x_2) + (1 - \alpha)f(x_1) - \frac{1}{2}c\alpha(1 - \alpha)\|x_2 - x_1\|^2. \quad (6.1.2)$$

It turns out that this “degree of non-affinity” is also measured by how much (6.1.1) deviates from equality: the next result is to be compared to Theorems IV.4.1.1(ii)-(iii) and IV.4.1.4, in a slightly different setting:  $f$  is now assumed convex but not differentiable.

**Theorem 6.1.2** *A necessary and sufficient condition for a convex function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  to be strongly convex (with modulus  $c > 0$ ) on a convex set  $C$  is that, for all  $x_1, x_2 \in C$ ,*

$$f(x_2) \geq f(x_1) + \langle s, x_2 - x_1 \rangle + \frac{1}{2}c\|x_2 - x_1\|^2 \quad \text{for all } s \in \partial f(x_1) \quad (6.1.3)$$

or equivalently

$$\langle s_2 - s_1, x_2 - x_1 \rangle \geq c\|x_2 - x_1\|^2 \quad \text{for all } s_i \in \partial f(x_i), i = 1, 2. \quad (6.1.4)$$

PROOF. For  $x_1, x_2$  given in  $C$  and  $\alpha \in ]0, 1[$ , we will use the notation

$$x^\alpha := \alpha x_2 + (1 - \alpha)x_1 = x_1 + \alpha(x_2 - x_1)$$

and we will prove (6.1.3)  $\Rightarrow$  (6.1.2)  $\Rightarrow$  (6.1.4)  $\Rightarrow$  (6.1.3).

[ $(6.1.3) \Rightarrow (6.1.2)$ ] Write (6.1.3) with  $x_1$  replaced by  $x^\alpha \in C$ : for  $s \in \partial f(x^\alpha)$ ,

$$f(x_2) \geq f(x^\alpha) + \langle s, x_2 - x^\alpha \rangle + \frac{1}{2}c\|x_2 - x^\alpha\|^2$$

or equivalently

$$f(x_2) \geq f(x^\alpha) + (1 - \alpha)\langle s, x_2 - x_1 \rangle + \frac{1}{2}c(1 - \alpha)^2\|x_2 - x_1\|^2.$$

Likewise,

$$f(x_1) \geq f(x^\alpha) + \alpha\langle s, x_1 - x_2 \rangle + \frac{1}{2}c\alpha^2\|x_1 - x_2\|^2.$$

Multiply these last two inequalities by  $\alpha$  and  $(1 - \alpha)$  respectively, and add to obtain

$$\alpha f(x_2) + (1 - \alpha)f(x_1) \geq f(x^\alpha) + \frac{1}{2}c\|x_2 - x_1\|^2[\alpha(1 - \alpha)^2 + (1 - \alpha)\alpha^2].$$

Then realize after simplification that this is just (6.1.2).

$[(6.1.2) \Rightarrow (6.1.4)]$  Write (6.1.2) as

$$\frac{f(x^\alpha) - f(x_1)}{\alpha} + \frac{1}{2}c(1 - \alpha)\|x_2 - x_1\|^2 \leq f(x_2) - f(x_1)$$

and let  $\alpha \downarrow 0$  to obtain

$$f'(x_1, x_2 - x_1) + \frac{1}{2}c\|x_2 - x_1\|^2 \leq f(x_2) - f(x_1)$$

which implies (6.1.3). Then, copying (6.1.3) with  $x_1$  and  $x_2$  interchanged and adding yields (6.1.4) directly.

$[(6.1.4) \Rightarrow (6.1.3)]$  Apply Theorem 2.3.4 to the one-dimensional convex function  $\mathbb{R} \ni \alpha \mapsto \varphi(\alpha) := f(x^\alpha)$ :

$$f(x_2) - f(x_1) = \varphi(1) - \varphi(0) = \int_0^1 \langle s^\alpha, x_2 - x_1 \rangle d\alpha \quad (6.1.5)$$

where  $s^\alpha \in \partial f(x^\alpha)$  for  $\alpha \in [0, 1]$ . Then take  $s_1$  arbitrary in  $\partial f(x_1)$  and apply (6.1.4):

$$\langle s^\alpha - s_1, x^\alpha - x_1 \rangle \geq c\|x^\alpha - x_1\|^2$$

i.e., using the value of  $x^\alpha$ ,

$$\alpha\langle s^\alpha, x_2 - x_1 \rangle \geq \alpha\langle s_1, x_2 - x_1 \rangle + c\alpha^2\|x_2 - x_1\|^2.$$

The result follows by using this inequality to minorize the integral in (6.1.5).  $\square$

Monotonicity properties of  $\partial f$  characterize strictly convex functions in just the same way as they do for strongly convex functions.

**Proposition 6.1.3** *A necessary and sufficient condition for a convex function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  to be strictly convex on a convex set  $C \subset \mathbb{R}^n$  is that, for all  $x_1, x_2 \in C$  with  $x_2 \neq x_1$ ,*

$$f(x_2) > f(x_1) + \langle s, x_2 - x_1 \rangle \quad \text{for all } s \in \partial f(x_1)$$

or equivalently

$$\langle s_2 - s_1, x_2 - x_1 \rangle > 0 \quad \text{for all } s_i \in \partial f(x_i), i = 1, 2.$$

PROOF. Copy the proof of Theorem 6.1.2 with  $c = 0$  and the relevant “ $\geq$ ”-signs replaced by strict inequalities. The only delicate point is in the  $[(6.1.2) \Rightarrow (6.1.4)]$ -stage: use monotonicity of the difference quotient.  $\square$

## 6.2 Continuity Properties of the Subdifferential

When  $f$  is a differentiable convex function, its gradient  $\nabla f$  is continuous, as a mapping from  $\mathbb{R}^n$  to  $\mathbb{R}^n$ . In the nondifferentiable case, this gradient becomes a set  $\partial f$  and our aim here is to study continuity properties of this set: to what extent can we say that  $\partial f(x)$  “varies continuously” with  $x$ , or with  $f$ ? We are therefore dealing with continuity properties of multifunctions, and we refer to §A.5 for the basic terminology.

We already know that  $\partial f(x)$  is compact convex for each  $x$ , and the next two results concern “global” properties.

**Proposition 6.2.1** *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be convex. The graph of its subdifferential mapping is closed in  $\mathbb{R}^n \times \mathbb{R}^n$ .*

PROOF. Let  $\{(x_k, s_k)\}$  be a sequence in  $\text{gr } \partial f$  converging to  $(x, s) \in \mathbb{R}^n \times \mathbb{R}^n$ . We must prove that  $(x, s) \in \text{gr } \partial f$ , which is easy. We have for all  $k$

$$f(y) \geq f(x_k) + \langle s_k, y - x_k \rangle \quad \text{for all } y \in \mathbb{R}^n;$$

pass to the limit on  $k$ , using continuity of  $f$  and of the scalar product.  $\square$

**Proposition 6.2.2** *The mapping  $\partial f$  is locally bounded, i.e. the image  $\partial f(B)$  of a bounded set  $B \subset \mathbb{R}^n$  is a bounded set in  $\mathbb{R}^n$ .*

PROOF. For arbitrary  $x$  in  $B$  and  $s \neq 0$  in  $\partial f(x)$ , the subgradient inequality implies in particular

$$f(x + s/\|s\|) \geq f(x) + \|s\|.$$

On the other hand,  $f$  is Lipschitz-continuous on the bounded set  $B + B(0, 1)$  (Theorem IV.3.1.2). Hence  $\|s\| \leq L$  for some  $L$ .  $\square$

**Remark 6.2.3** Combining these two results, we obtain a bit more than compact-valuedness of  $\partial f$ , namely: the image by  $\partial f$  of a *compact* set is compact. In fact, for  $\{x_k\}$  in a compact set, with a subsequence  $\{x_{k'}\}$ , say, converging to  $x$ , take  $s_k \in \partial f(x_k)$  and extract the subsequence  $\{s_{k'}\}$ . From Proposition 6.2.2, a subsequence of  $\{s_{k'}\}$  converges to some  $s \in \mathbb{R}^n$ ; from Proposition 6.2.1,  $s \in \partial f(x)$ . As another consequence, we obtain for example: the image by  $\partial f$  of a *compact connected* set is compact connected.

On the other hand, the image by  $\partial f$  of a *convex* set is certainly *not* convex (except for  $n = 1$ , where convexity and connectedness coincide): take for  $f$  the  $\ell_1$ -norm on  $\mathbb{R}^2$ ; the image by  $\partial f$  of the unit simplex  $\Delta_2$  is the union of two segments which are not collinear.

Concerning the graph of  $\partial f$ , the same type of results hold: if  $K \subset \mathbb{R}^n$  is compact connected, the set

$$\{(x, s) \in \mathbb{R}^n \times \mathbb{R}^n : x \in K, s \in \partial f(x)\}$$

is compact connected in  $\mathbb{R}^n \times \mathbb{R}^n$ . Also, it is a “skinny” set (see again Fig. I.4.1.1) because  $\partial f(x)$  is a singleton almost everywhere (Theorem IV.4.2.3).  $\square$

Thanks to local boundedness, our mapping  $\partial f$  takes its values in a compact set when the argument  $x$  itself varies in a compact set; the “nice” form (A.5.2) of outer and inner semi-continuity can then be used.

**Theorem 6.2.4** *The subdifferential mapping of a convex function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is outer semi-continuous at any  $x \in \mathbb{R}^n$ , i.e.*

$$\forall \varepsilon > 0, \exists \delta > 0 : y \in B(x, \delta) \implies \partial f(y) \subset \partial f(x) + B(0, \varepsilon). \quad (6.2.1)$$

PROOF. Assume for contradiction that, at some  $x$ , there are  $\varepsilon > 0$  and a sequence  $\{(x_k, s_k)\}$  with

$$\begin{aligned} x_k &\rightarrow x \quad \text{for } k \rightarrow \infty & \text{and} \\ s_k \in \partial f(x_k), \quad s_k &\notin \partial f(x) + B(0, \varepsilon) \quad \text{for } k = 1, 2, \dots \end{aligned} \quad (6.2.2)$$

The bounded  $\{s_k\}$  (Proposition 6.2.2) has a subsequence converging to  $s$ , which is in  $\partial f(x)$  (Proposition 6.2.1). This is a contradiction since (6.2.2) implies

$$s \notin \partial f(x) + B(0, \frac{1}{2}\varepsilon).$$

□

In terms of directional derivatives, we recover a natural result, if we remember that  $f'(\cdot, d)$  is an infimum of continuous functions  $[f(\cdot + td) - f(\cdot)]/t$  over  $t > 0$ :

**Corollary 6.2.5** *For  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  convex, the function  $f'(\cdot, d)$  is upper semi-continuous: at all  $x \in \mathbb{R}^n$ ,*

$$f'(x, d) = \limsup_{y \rightarrow x} f'(y, d) \quad \text{for all } d \in \mathbb{R}^n.$$

PROOF. Use Theorem 6.2.4, in conjunction with Proposition V.3.3.9. □

**Remark 6.2.6** If  $f$  is differentiable at  $x$ , then Theorem 6.2.4 reads as follows: all the subgradients at  $y$  tend to  $\nabla f(x)$  when  $y$  tends to  $x$ . The inner semi-continuity then follows:  $\partial f$  is actually continuous at  $x$ . In particular, if  $f$  is differentiable on an open set  $\Omega$ , then it is continuously differentiable on  $\Omega$ .

In the general case, however, inner semi-continuity is hopeless: for  $n = 1$  and  $f(x) := |x|$ ,  $\partial f$  is not inner semi-continuous at 0:  $\partial f(0) = [-1, +1]$  is much larger than, say,  $\partial f(x) = \{1\}$  when  $x > 0$ . □

All the previous results concerned the behaviour of  $\partial f(x)$  as varying with  $x$ . This behaviour is essentially the same when  $f$  varies as well.

**Theorem 6.2.7** *Let  $\{f_k\}$  be a sequence of (finite) convex functions converging pointwise to  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  and let  $\{x_k\}$  converge to  $x \in \mathbb{R}^n$ . For any  $\varepsilon > 0$ ,*

$$\partial f_k(x_k) \subset \partial f(x) + B(0, \varepsilon) \quad \text{for } k \text{ large enough.}$$

PROOF. Let  $\varepsilon > 0$  be given. Recall (Theorem IV.3.1.5) that the pointwise convergence of  $\{f_k\}$  to  $f$  implies its uniform convergence on every compact set of  $\mathbb{R}^n$ .

First, we establish boundedness: for  $s_k \neq 0$  arbitrary in  $\partial f_k(x_k)$ , we have

$$f_k(x_k + s_k/\|s_k\|) \geq f_k(x_k) + \|s_k\|.$$

The uniform convergence of  $\{f_k\}$  to  $f$  on  $B(x, 2)$  implies for  $k$  large enough

$$\|s_k\| \leq f(x_k + s_k/\|s_k\|) - f(x_k) + \varepsilon,$$

and the Lipschitz property of  $f$  on  $B(x, 2)$  ensures that  $\{s_k\}$  is bounded.

Now suppose for contradiction that, for some infinite subsequence, there is some  $s_k \in \partial f_k(x_k)$  which is not in  $\partial f(x) + B(0, \varepsilon)$ . Any cluster point of this  $\{s_k\}$  – and there is at least one – is out of  $\partial f(x) + B(0, 1/2\varepsilon)$ . Yet, with  $y$  arbitrary in  $\mathbb{R}^n$ , write

$$f_k(y) \geq f_k(x_k) + \langle s_k, y - x_k \rangle$$

and pass to the limit (on a further subsequence such that  $s_k \rightarrow s$ ): pointwise [resp. uniform] convergence of  $\{f_k\}$  to  $f$  at  $y$  [resp. around  $x$ ], and continuity of the scalar product give

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

Because  $y$  was arbitrary, we obtain the contradiction  $s \in \partial f(x)$ .  $\square$

The differentiable case is worth mentioning:

**Corollary 6.2.8** *Let  $\{f_k\}$  be a sequence of (finite) differentiable convex functions converging pointwise to the differentiable  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . Then  $\nabla f_k$  converges to  $\nabla f$  uniformly on every compact set of  $\mathbb{R}^n$ .*

PROOF. Take  $S$  compact and suppose for contradiction that there exist  $\varepsilon > 0$ ,  $\{x_k\} \subset S$  such that

$$\|\nabla f_k(x_k) - \nabla f(x_k)\| > \varepsilon \quad \text{for } k = 1, 2, \dots$$

Extracting a subsequence if necessary, we may suppose  $x_k \rightarrow x \in S$ ; Theorem 6.2.7 assures that  $\{\nabla f_k(x_k)\}$  and  $\{\nabla f(x_k)\}$  both converge to  $\nabla f(x)$ , implying  $0 \geq \varepsilon$ .  $\square$

### 6.3 Subdifferentials and Limits of Gradients

One of the main results of the previous section was the outer semi-continuity of the subdifferential: (6.2.1) just means that this latter set contains all the possible limits of subgradients calculated at all neighboring points.

The question that we consider in this section is in a sense the converse: to what extent can the whole subdifferential be built up from *limits* of subgradients at neighboring points? In other words: we are given  $x \in \mathbb{R}^n$  and we want to construct sequences  $\{(y_k, s_k)\} \subset \text{gr } \partial f$  so that the limits of  $\{s_k\}$  make up the entire  $\partial f(x)$ . Of course, we are not too interested in the trivial case where  $y_k \equiv x$ ; we will actually consider two special kinds of sequences  $\{y_k\}$ .

**(a) Sequences of Differentiability Points** First, consider sequences  $\{y_k\}$  such that  $f$  is differentiable at each  $y_k$ . Recall from Theorem IV.4.2.3 that  $f$  is differentiable except possibly on a set of measure zero; call it  $\Delta^c$ , i.e.

$$y \in \Delta \iff \partial f(y) = \{\nabla f(y)\}.$$

Thus, even if our given  $x$  is not in  $\Delta$ , we can construct a sequence  $\{y_k\} \subset \Delta$  with  $y_k \rightarrow x$ . The corresponding sequence  $\{\nabla f(y_k)\}$  is bounded (by a Lipschitz constant of  $f$  around  $x$ ), so we can extract some cluster point; according to §6.2, any such cluster point is in  $\partial f(x)$ . Then we ask the question: how much of  $\partial f(x)$  do we cover with all the possible subgradients obtained with this limiting process?

Example 3.4 can be used to illustrate the construction above: with the  $x$  of the right part of Fig. 3.1, we let  $\{y_k\}$  be any sequence tending to  $x$  and keeping away from the kinky line where  $f_1 = f_2$ . For example, with the  $d$  of the picture, we can take

$$y_k := x + \frac{(-1)^k}{k} d,$$

in which case the corresponding sequence  $\{\nabla f(y_k)\}$  has two cluster points  $s_1$  and  $s_2$  – and our set of limits is complete: no other sequence of gradients can produce another limit. Observe in this example that  $\partial f(x)$  is the convex hull of the cluster points  $s_1$  and  $s_2$  thus obtained. We will show that this is always the case.

So we set

$$\gamma f(x) := \{s : \exists \{y_k\} \subset \Delta \text{ with } y_k \rightarrow x, \nabla f(y_k) \rightarrow s\}. \quad (6.3.1)$$

It is rather clear that  $\gamma f(x)$  is bounded, and also that it is closed (as a “limit of limits”); its convex hull is therefore compact (Theorem III.1.4.3) and, by Theorem 6.2.4,

$$\gamma f(x) \subset \text{co } \gamma f(x) \subset \partial f(x). \quad (6.3.2)$$

The next result establishes the converse inclusion.

**Theorem 6.3.1** *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be convex. With the notation (6.3.1),*

$$\partial f(x) = \text{co } \gamma f(x) \quad \text{for all } x \in \mathbb{R}^n. \quad (6.3.3)$$

PROOF. In view of (6.3.2), we only have to prove that

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

where the support function of  $\gamma f(x)$  is obtained from (6.3.1):

$$\sigma_{\gamma f(x)}(d) = \limsup \{\langle \nabla f(y), d \rangle : y \rightarrow x, y \in \Delta\}. \quad (6.3.4)$$

Suppose that, for some  $\varepsilon > 0$  and (normalized)  $d$ , it holds that

$$\sigma_{\gamma f(x)}(d) < f'(x, d) - \varepsilon.$$

In view of the formulation (6.3.4), this means that, for some  $\delta > 0$ ,

$$\langle \nabla f(x'), d \rangle \leq f'(x, d) - \frac{1}{2}\varepsilon \quad \text{for all } x' \in B(x, \delta) \cap \Delta.$$

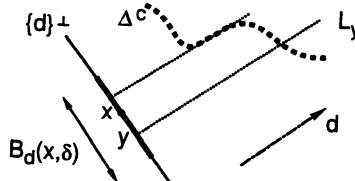
Consider the following set (see Fig. 6.3.1):

$$B_d(x, \delta) := \{y \in B(x, \delta) : \langle d, y \rangle = 0\}$$

and, for each  $y \in B_d(x, \delta)$ , denote by  $L_y := y + \mathbb{R}d$  the line passing through  $y$  and parallel to  $d$ . According to Fubini's Theorem, the 1-dimensional measure of  $L_y \cap \Delta^c$  is zero for almost all  $y \in B_d(x, \delta)$  (equipped with its  $(n-1)$ -dimensional measure). Take such a  $y$ , so that  $f$  is differentiable at almost all points of the form  $y + td$ ; then write

$$f'(y, d) \leq \frac{f(y + td) - f(y)}{t} = \frac{1}{t} \int_0^t \langle \nabla f(y + \alpha d), d \rangle d\alpha \leq f'(x, d) - \frac{\varepsilon}{2},$$

which contradicts the upper semi-continuity of  $f'(\cdot, d)$  at  $x$  (Corollary 6.2.5).  $\square$



**Fig. 6.3.1.** Meeting the set of kinks of a convex function

In summary, the subdifferential can be reconstructed as the convex hull of all possible limits of gradients at points  $y_k$  tending to  $x$ . In addition to 1.1.4, 1.2.1 and 1.3.1, a fourth possible definition of  $\partial f(x)$  is (6.3.3).

**Remark 6.3.2** As a consequence of the above characterization, we indicate a practical trick to compute subgradients: pretend that the function is differentiable and proceed as usual in differential calculus. This “rule of thumb” can be quite useful in some situations.

For an illustration, consider the example of §5.1. Once the largest eigenvalue  $\lambda$  is computed, together with an associated eigenvector  $u$ , just pretend that  $\lambda$  has multiplicity 1 and differentiate formally the equation  $Mu = \lambda u$ . A differential  $dM$  induces the differentials  $du$  and  $d\lambda$  satisfying

$$Md u + dM u = \lambda du + d\lambda u.$$

We need to eliminate  $du$ ; for this, premultiply by  $u^\top$  and use symmetry of  $M$ , i.e.  $u^\top M = \lambda u^\top$ :

$$\lambda u^\top du + u^\top dM u = \lambda u^\top du + d\lambda u^\top u.$$

Observing that  $u^\top dM u = uu^\top dM$ ,  $d\lambda$  is obtained as a linear form of  $dM$ :

$$d\lambda = \frac{uu^\top}{u^\top u} dM = S dM.$$

Moral: if  $\lambda$  is differentiable at  $M$ , its gradient is the matrix  $S$ ; if not, we find the expression of a subgradient (depending on the particular  $u$ ).

This trick is only heuristic, however, and Remark 4.1.2 warns us against it: for  $n = 1$ , take  $f = f_1 - f_2$  with  $f_1(x) = f_2(x) = |x|$ ; if we “differentiate” naively  $f_1$  and  $f_2$  at 0 and subtract the “derivatives” thus obtained, there are 2 chances out of 3 of ending up with a wrong result.  $\square$

**Example 6.3.3** Let  $d_C$  be the distance-function to a nonempty closed convex set  $C$ . From Example 3.3, we know several things: the kinks of  $d_C$  form the set  $\Delta^c = \text{bd } C$ ; for  $x \in \text{int } C$ ,  $\nabla d_C(x) = 0$ ; and

$$\nabla d_C(x) = \frac{x - p_C(x)}{\|x - p_C(x)\|} \quad \text{for } x \notin C. \quad (6.3.5)$$

Now take  $x_0 \in \text{bd } C$ ; we give some indications to construct  $\partial d_C(x_0)$  via (6.3.3) (draw a picture).

- First,  $\gamma d_C(x_0)$  contains all the limits of vectors of the form (6.3.5), for  $x \rightarrow x_0$ ,  $x \notin C$ . These can be seen to make up the intersection of the normal cone  $N_C(x_0)$  with the unit sphere (technically, the multifunction  $x \mapsto \partial d_C(x)$  is outer semi-continuous).
- If  $\text{int } C \neq \emptyset$ , append  $\{0\}$  to this set; the description of  $\gamma d_C(x_0)$  is now complete.
- As seen in Example 3.3, the convex hull of the result must be the truncated cone  $N_C(x_0) \cap B(0, 1)$ . This is rather clear in the second case, when  $0 \in \gamma d_C(x_0)$ ; but it is also true even if  $\text{int } C = \emptyset$ : in fact  $N_C(x_0)$  contains the subspace orthogonal to  $\text{aff } C$ , which in this case contains two opposite vectors of norm 1.  $\square$

**(b) Directional Sequences** We consider now a second type of sequences: those of the form  $x + t_k d$ , for fixed normalized  $d$  and  $t_k \downarrow 0$ . We start with a fairly easy but important lemma, which supplements the closedness result 6.2.1.

**Lemma 6.3.4** *Let  $x$  and  $d$  with  $\|d\| = 1$  be given in  $\mathbb{R}^n$ . For any sequence  $\{(t_k, s_k)\} \subset \mathbb{R}_+^+ \times \mathbb{R}^n$  satisfying*

$$t_k \downarrow 0 \quad \text{and} \quad s_k \in \partial f(x + t_k d) \text{ for } k = 1, 2, \dots$$

*and any cluster point  $s$  of  $\{s_k\}$ , there holds*

$$s \in \partial f(x) \quad \text{and} \quad \langle s, d \rangle = f'(x, d).$$

PROOF. The first property comes from the results in §6.2. For the second, use the monotonicity of  $\partial f$ :

$$0 \leq \langle s_k - s', x + t_k d - x \rangle = t_k \langle s_k - s', d \rangle \quad \text{for all } s' \in \partial f(x).$$

Divide by  $t_k > 0$  and pass to the limit to get  $f'(x, d) \leq \langle s, d \rangle$ . The converse inequality being trivial, the proof is complete.  $\square$

In other words, taking a limit of subgradients from a directional sequence amounts to taking a subgradient which is not arbitrary, but which lies in a designated face of  $\partial f(x)$ : the face  $F_{\partial f(x)}(d)$  exposed by the direction  $d$  in question. When this direction describes the unit sphere, each exposed face of  $\partial f(x)$  is visited (Proposition V.3.1.5). We thus obtain a second set of subgradients, analogously to the way  $\gamma f(x)$  was constructed by (6.3.1).

More precisely, suppose that we have a process (call it  $\Pi$ ) which, given  $x$  and the normalized  $d$ , does the following:

- form a directional sequence  $y_k = x + t_k d$  tending to  $x$ ;
- for each  $k$ , select a subgradient  $s_k \in \partial f(y_k)$ ;
- take a cluster point  $s$  of  $\{s_k\}$ .

We call  $s(d) \in \partial f(x)$  the subgradient thus obtained – a notation  $s_{\Pi}(d)$  would be more correct, to emphasize the dependence of  $s(d)$  on the particular process used.

**Remark 6.3.5** In view of Lemma 6.3.4, another process would do the same kind of job, namely:

- Maximize  $\langle s, d \rangle$  over  $s \in \partial f(x)$  to obtain some solution  $s(d)$  – or  $s_{\Pi}(d)$ .

We took the trouble to describe the more complicated process  $\Pi$  above because the concept of directional sequences is important for minimization algorithms to be studied later, starting from Chap. IX.  $\square$

Now we form the set of all outputs of the above process  $\Pi$  (whatever it may be), for all directions  $d$ :

$$\delta f(x) := \{s(d) : d \in \mathbb{R}^n, \|d\| = 1\} \quad [\text{or } \delta_{\Pi} f(x) := \cup_{\|d\|=1} s_{\Pi}(d)].$$

Once again,  $\delta_{\Pi} f(x)$  is a compact set included in  $\partial f(x)$  and there holds

$$\delta_{\Pi} f(x) \subset \text{co } \delta_{\Pi} f(x) \subset \partial f(x).$$

It turns out that Theorem 6.3.1 can be reproduced:

**Theorem 6.3.6** *No matter how the process  $\Pi$  is chosen to generate each cluster point  $s_{\Pi}(d)$ , it holds that*

$$\partial f(x) = \text{co } \delta_{\Pi} f(x). \quad (6.3.6)$$

PROOF. We have to prove only the “ $\subset$ ”-inclusion in (6.3.6). Use Lemma 6.3.4: for each  $d$  of norm 1, the  $s(d)$  generated by the process satisfies

$$\sigma_{\partial f(x)}(d) = f'(x, d) = \langle s(d), d \rangle \leq \sigma_{\delta_{\Pi} f(x)}(d).$$

$\square$

A fifth possible definition of the subdifferential of  $f$  at  $x$  is thus given by (6.3.6).

**Remark 6.3.7** As an application, consider the following problem: given a (finite) sublinear function  $\sigma$ , how can we construct its supported set  $\partial\sigma(0)$ ? Answer: differentiate  $\sigma$  wherever possible;  $\partial\sigma(0)$  is then the closed convex hull of the collection of gradients thus obtained.

In fact, if the gradient  $\nabla\sigma(d)$  exists, it exists (and stays the same) all along the ray  $\mathbb{R}_*^+ d$ : we are therefore constructing  $\partial\sigma(0)$ ; and we can limit ourselves to computing  $\nabla\sigma$  on the unit sphere.

For example, the  $\ell_1$ -norm  $|x|_1 = \sum_{i=1}^n |\xi^i|$  can be differentiated whenever no  $\xi^i$  is 0; the resulting gradients are the vectors whose components are  $\pm 1$ . Their convex hull is  $\partial|\cdot|_1(0)$  and we refer to §V.3.2 for the various interpretations that this set can be given, in terms of polarity, duality, sublevel-sets, gauges, etc.

The linear function  $\langle \cdot, d \rangle$  being maximized in  $\partial\sigma(0)$  on the face exposed by  $d$ , we write

$$\sigma(d) = \langle \partial\sigma(d), d \rangle.$$

When  $\partial\sigma(d)$  is the singleton  $\{\nabla\sigma(d)\}$ , this is known as *Euler's relation* associated with the positively homogeneous function  $\sigma$ . Remember also the geometric construction:  $d \neq 0$  defines the hyperplane

$$H_{d,\sigma(d)} = \{s \in \mathbb{R}^n : \langle s, d \rangle = \sigma(d)\},$$

which remains fixed when  $d$  is replaced by  $\kappa d$ ,  $\kappa > 0$ , and which envelopes  $\partial\sigma(0)$  when  $d$  describes  $\mathbb{R}^n \setminus \{0\}$ .  $\square$

To conclude this chapter, we point out a rather important property of convex functions revealed by Lemma 6.3.4: to expose a face in some subdifferential  $\partial f(x)$  along a direction  $d$ , it suffices to know the trace of  $f$  along  $x + \mathbb{R}^+ d$ . By contrast, exposing a face in an abstract closed convex set  $C$  requires the maximization of a linear function over  $C$  – which somehow implies a full knowledge of  $C$ .

With this in mind, a geometric interpretation of the process in §1.4 can be given. When we compute  $\sigma_1 = \sigma'_0(e_1, \cdot)$ , we simply expose by  $e_1$  a face of  $\partial\sigma_0(0) = \partial f(x)$ :

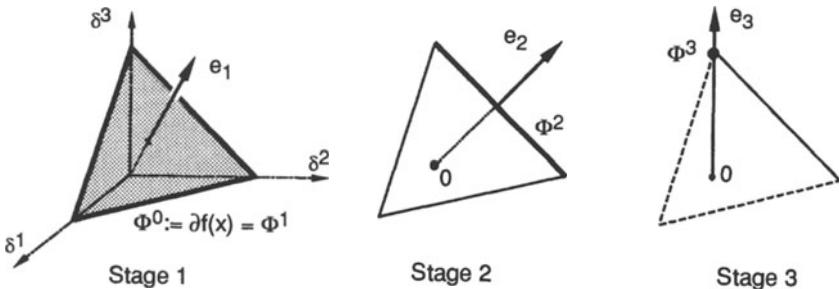
$$\Phi^1 := \partial\sigma_1(0) = F_{\partial\sigma_0(0)}(e_1) = F_{\partial f(x)}(e_1).$$

Needless to say, the breadth (in the sense of Definition V.2.1.4) of  $\Phi^1$  along  $e_1$  is zero – another way of saying that  $\sigma_0$  is differentiable at  $e_1$  in the subspace  $\mathbb{R}e_1$ . As a result, the dimension of  $\Phi^1$  is at most  $n - 1$ . Then, we extract recursively from  $\Phi^{k-1}$  the face exposed by  $e_k$ :

$$\Phi^k := \partial\sigma_k(0) = F_{\partial\sigma_{k-1}(0)}(e_k) = F_{\Phi^{k-1}}(e_k).$$

At each stage,  $\dim \Phi^k$  thus loses one unit. We end up with a face of  $\Phi^{n-1}$  which is certainly of dimension 0.

Remembering Remark III.2.4.4, we see that the subgradient  $\Phi^n$  is not quite arbitrary: it is a subface (more precisely a *vertex*) of each  $\Phi^k$ , in particular of the original set  $\partial f(x)$ . Figure 6.3.2 illustrates the process applied to Example 1.4.3 (use Remark 6.3.7 to realize that  $\partial\sigma_0(0)$  is the unit simplex).



**Fig. 6.3.2.** Exposing a vertex in a subdifferential

## VII. Constrained Convex Minimization Problems: Minimality Conditions, Elements of Duality Theory

**Prerequisites.** Subdifferentials of finite convex functions (Chap. VI); tangent and normal cones to convex sets (Chap. III).

**Introduction.** The basic *convex minimization problem* is that of finding some  $\bar{x} \in C$  such that

$$f(\bar{x}) = \inf \{f(x) : x \in C\}, \quad (0.1)$$

where  $C \subset \mathbb{R}^n$  and  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  are a closed convex set and a (finite-valued) convex function respectively; the points in  $C$  are called *feasible*; a solution  $\bar{x}$  must therefore satisfy two properties: feasibility, and optimality.

In this chapter, we are mainly interested in *characterizing* a solution in terms of the data of the problem: the constraint-set  $C$  and the objective function  $f$ . This study is indeed preliminary to an actual resolution of (0.1), possibly an approximate one via the construction of a minimizing sequence. An exact solution can be constructed in simple cases, such as *quadratic programming*, where  $f$  is quadratic (convex) and  $C$  is a closed convex polyhedron.

The question of conditions for a candidate  $\bar{x}$  to be a solution was already evoked in §II.1.1: starting just from the definition of a minimum,

$$f(\bar{x}) \leq f(x) \quad \text{for all } x \in C \quad (\text{and of course: } \bar{x} \in C),$$

we want to derive more useful conditions, using the “tangential elements” of the data, namely the directional derivative of  $f$  and the tangent cone to  $C$ . Dually, we will also use the subdifferential of  $f$  and the normal cone to  $C$ . The situation is here more involved than in Chap. II because we have constraints, but on the other hand convexity makes it simpler to some extent. In particular, necessary *and* sufficient conditions are available, while sufficiency is usually out of reach in the nonconvex case.

Depending on the properties and information concerning the data, various approaches can be relevant, each based on appropriate techniques.

- When  $C$  is not specified, no wonder that the only available conditions are “abstract”, or “formal”: they involve the tangent cone  $T_C$  and normal cone  $N_C$  and nothing more. This will be the subject of §1.
- When  $C$  is described more explicitly (§2), these cones can themselves be characterized more explicitly. The most important case is a representation of the closed convex  $C$  by *constraints*:

$$\begin{cases} \langle a_i, x \rangle = b_i & \text{for } i = 1, \dots, m, \\ c_j(x) \leq 0 & \text{for } j = 1, \dots, p. \end{cases} \quad (0.2)$$

where the  $c_j$ 's are finite-valued convex functions, so as to make  $C$  convex (equality constraints are taken to be affine for the same reason). When expressing the minimality conditions, the subdifferentials  $\partial c_j$  and the gradients  $a_i$  will certainly show up, as will  $\partial f$ . This can be done in two different ways: one is to expand the expression of  $T_C$  and  $N_C$  in the “formal” minimality conditions of §1; the other is to tackle the problem directly, linearizing the functions ( $f$  and)  $c_j$ .

We limit our study to finite-valued functions. One reason is that we make extensive use of subdifferentials, studied in Chap. VI in this framework only; and also, this is sufficient to capture the essential features of any convex minimization problem. Just as in Chap. VI, it would of course suffice to assume that the candidate  $\bar{x}$  to minimality is interior to the domain of the functions involved.

In the case of a description by (0.2), a solution  $\bar{x}$  of (0.1) is essentially characterized by the existence of  $m + p$  numbers  $\lambda_1, \dots, \lambda_m, \mu_1, \dots, \mu_p$  – the *multipliers* – satisfying

$$0 \in \partial f(\bar{x}) + \sum_{i=1}^m \lambda_i a_i + \sum_{j=1}^p \mu_j \partial c_j(\bar{x}) \text{ and} \\ \text{for } j = 1, \dots, p, \quad \mu_j \geq 0 \text{ and } \mu_j = 0 \text{ if } c_j(\bar{x}) < 0. \quad (0.3)$$

There are infinitely many ways of representing a set via constraints as in (0.2), and it turns out that a characterization like (0.3) cannot be expected to hold in *all* cases. Indeed, the data have to satisfy some assumption: a *constraint qualification* condition.

In view of the calculus rule VI.4.1.1, (0.3) displays the subdifferential of a certain function: the *Lagrange function*

$$L(x, \lambda, \mu) := f(x) + \sum_{i=1}^m \lambda_i (\langle a_i, x \rangle - b_i) + \sum_{j=1}^p \mu_j c_j(x)$$

which is central in all this theory. Its role will be the subject of §3. Indeed,  $\bar{x}$  minimizes  $L(\cdot, \lambda, \mu)$  for some  $(\lambda, \mu) \in \mathbb{R}^m \times (\mathbb{R}^+)^p$  and we will see that the couple  $(\lambda, \mu)$  maximizes  $L(\bar{x}, \cdot, \cdot)$ . This observation motivates our Section 4, in which we give some account of *duality* and *saddle-point* problems.

Unless otherwise specified, we postulate that our original constrained minimization problem (0.1) does have an optimal solution. Thus, the general framework throughout this chapter is as follows: we have a convex function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , a nonempty closed convex set  $C \subset \mathbb{R}^n$ , and  $f$  assumes its minimum over  $C$  at some  $\bar{x} \in C$ . We will sometimes denote by

$$S := \{\bar{x} \in C : f(\bar{x}) \leq f(x) \text{ for all } x \in C\} \neq \emptyset$$

the solution-set of (0.1).

## 1 Abstract Minimality Conditions

We start with the “abstract” convex minimization problem

$$\inf \{f(x) : x \in C\}, \quad (1.0.1)$$

for which we characterize a solution (assumed to exist) in terms of the data  $(C, f)$  alone and their “first-order elements”  $(T_C, f'; N_C, \partial f)$ .

First of all, we make sure that, just as in the unconstrained case, there can be no ambiguity in the definition of a minimum:

**Lemma 1.0.1** *Any local minimum of  $f$  on  $C$  is a global minimum. Furthermore, the set of minima is a closed convex subset of  $C$ .*

PROOF. Suppose that, for some  $r > 0$ ,

$$f(\bar{x}) \leq f(x) \quad \text{for all } x \in B(\bar{x}, r) \cap C.$$

Take  $x \in C$ ,  $x \notin B(\bar{x}, r)$  and set  $t := r/\|x - \bar{x}\|$ ,  $x_t := (1 - t)\bar{x} + tx$ . Clearly,  $0 < t < 1$  and  $x_t \in B(\bar{x}, r) \cap C$ ; by assumption and using the convexity of  $f$ ,

$$f(\bar{x}) \leq f(x_t) \leq (1 - t)f(\bar{x}) + tf(x).$$

Hence,  $tf(\bar{x}) \leq tf(x)$ , so  $\bar{x}$  is a global minimum of  $f$  on  $C$ .

On the other hand, the solution-set

$$C \cap \{x \in \mathbb{R}^n : f(x) \leq f(\bar{x})\}$$

is the intersection of two closed convex sets, and is therefore closed and convex.  $\square$

## 1.1 A Geometric Characterization

**Theorem 1.1.1** *With  $f$  and  $C$  as above, the following statements are equivalent when  $\bar{x} \in C$ :*

- (i)  $\bar{x}$  minimizes  $f$  over  $C$ ;
- (ii)  $f'(\bar{x}, y - \bar{x}) \geq 0$  for all  $y \in C$ ;
- (ii')  $f'(\bar{x}, d) \geq 0$  for all  $d \in T_C(\bar{x})$ ;
- (iii)  $0 \in \partial f(\bar{x}) + N_C(\bar{x})$ .

PROOF. [(i)  $\Rightarrow$  (ii)  $\Rightarrow$  (ii') ] Pick an arbitrary  $y \in C$ : by convexity,  $\bar{x} + t(y - \bar{x}) \in C$  for all  $t \in [0, 1]$ . If (i) holds, (ii) follows by letting  $t \downarrow 0$  in

$$\frac{f(\bar{x} + t(y - \bar{x})) - f(\bar{x})}{t} \geq 0, \quad \text{valid for all } t \in ]0, 1].$$

Setting  $d := y - \bar{x}$  and using positive homogeneity, we thus have  $f'(\bar{x}, d) \geq 0$  for all  $d$  in the cone  $\mathbb{R}^+(C - \bar{x})$ , whose closure is  $T_C(\bar{x})$  (Proposition III.5.2.1). Then (ii') follows from the continuity of the finite convex function  $f'(\bar{x}, \cdot)$  (Remark VI.1.1.3).

[(ii')  $\Rightarrow$  (i)] For arbitrary  $y \in C$ , we certainly have  $y - \bar{x} \in T_C(\bar{x})$ , hence (ii') implies

$$0 \leq f'(\bar{x}, y - \bar{x}) \leq f(y) - f(\bar{x})$$

(where the second inequality comes from the Definition VI.1.1.1 of the directional derivative of  $f$ ) and (i) is established.

[ $(ii') \Leftrightarrow (iii)$ ] Because  $f'(\bar{x}, \cdot)$  is finite everywhere, (ii') can be rewritten as follows:

$$f'(\bar{x}, d) + I_{T_C}(\bar{x})(d) \geq 0 \quad \text{for all } d \in \mathbb{R}^n.$$

The indicator  $I$  of the closed convex cone  $T_C(\bar{x})$  is the support  $\sigma$  of its polar cone (see Example V.2.3.1), which is  $N_C(\bar{x})$  (by Proposition III.5.2.4); also,  $f'(\bar{x}, \cdot)$  is the support of  $\partial f(\bar{x})$  (by Definition VI.1.1.4); using the calculus rule V.3.3.3(i) on the sum of support functions, we therefore obtain

$$(ii') \iff 0 \leq \sigma_{\partial f}(\bar{x}) + \sigma_{N_C}(\bar{x}) = \sigma_{\partial f(\bar{x}) + N_C(\bar{x})}.$$

Recall that the sum of the compact set  $\partial f(\bar{x})$  and of the closed convex set  $N_C(\bar{x})$  is a closed convex set: the above inequality is just (iii) in terms of support functions, thanks to Theorem V.2.2.2.  $\square$

When (1.0.1) is unconstrained,  $C = \mathbb{R}^n$  and, for all  $x$ ,  $T_C(x) = \mathbb{R}^n$ ,  $N_C(x) = \{0\}$ ; the above minimality conditions reduce to those of Theorem VI.2.2.1. The other extreme case  $C = \{x\}$  presents no interest! Of course, it is a privilege of convexity that (ii) – (iii) are *sufficient* for minimality, just because  $f'(x, \cdot)$  underestimates  $f(x + \cdot) - f(x)$  [i.e.  $\text{epi } f'(x, \cdot)$  contains  $\text{epi } f - \{(x, f(x))\}$ ] and  $T_C(x)$  overestimates [contains]  $C - \{x\}$ . This confirms Lemma 1.0.1: a *local* minimality condition is actually *global*.

**Remark 1.1.2** While  $f'(x, \cdot)$  is the tangential approximation of  $f$  near  $x$ ,  $T_C(x)$  is the set of tangent directions to  $C$  at  $x$ . The minimization problem

$$\inf_d \{f'(\bar{x}, d) : d \in T_C(\bar{x})\} \tag{1.1.1}$$

is the “tangent problem to (1.0.1)” at  $\bar{x}$  and (ii') says that its infimal value is nonnegative (in fact exactly 0, achieved at  $d = 0$ ). The negation of (ii') is that the infimal value is  $-\infty$ .

The tangent problem can be rephrased via the change of variable  $d = y - \bar{x}$ : consider the first-order approximation of  $f$  near  $\bar{x}$

$$y \mapsto \varphi_{\bar{x}}(y) := f(\bar{x}) + f'(\bar{x}, y - \bar{x}) \quad [= f(y) + o(\|y - \bar{x}\|)]$$

and plug the translation  $(0, 0) \mapsto (\bar{x}, f(\bar{x})) \in \mathbb{R}^n$  into (1.1.1), which becomes

$$\inf_y \{\varphi_{\bar{x}}(y) : y - \bar{x} \in T_C(\bar{x})\}. \tag{1.1.2}$$

If  $\bar{x}$  minimizes  $f$  over  $C$ , then  $y = \bar{x}$  solves (1.1.2) – possibly together with other solutions. Conversely, if  $\bar{x}$  does not minimize  $f$  over  $C$ , then (1.1.2) has no solution “at finite distance”. Observe that  $\varphi_{\bar{x}}$ , and also  $\{\bar{x}\} + T_C(\bar{x})$ , could be replaced by more accurate approximations of  $f$  and  $C$ , using a second-order approximation of  $f$ , for example. The essence of the result would not be changed: a solution of (1.0.1) would again be characterized as solving (1.1.2), modified accordingly. The equivalent condition (ii) does just this, replacing  $\{\bar{x}\} + T_C(\bar{x})$  by  $C$  itself, i.e. no approximation at all.

Still another way of reading (ii') is that  $f$  is locally increasing along each element of the set  $\mathbb{R}^+(C - \bar{x})$  of feasible directions for  $C$  at  $\bar{x}$ ; and this property is conserved when taking limits of such feasible directions, i.e. passing to  $T_C(\bar{x})$ .  $\square$

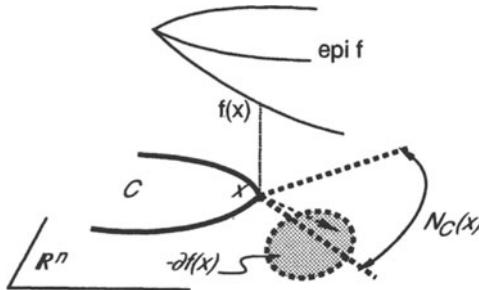
Condition (iii) appears as a dual formulation of (ii'), insofar as subgradients and normals are in the dual of  $\mathbb{R}^n$ . It can also be written

$$-\partial f(\bar{x}) \cap N_C(\bar{x}) \neq \emptyset, \quad (1.1.3)$$

which lends itself to a translation in plain words: there is *some subgradient* of  $f$  at  $\bar{x}$  whose opposite is *normal* to  $C$  at  $\bar{x}$ . If  $f$  is differentiable at  $\bar{x}$ , this means that  $-\nabla f(\bar{x})$  points inside the normal cone:  $-\nabla f(\bar{x})$  makes an obtuse angle [or  $\nabla f(\bar{x})$  makes an acute angle] with all feasible directions for  $C$  at  $\bar{x}$ ; see Fig. 1.1.1. When  $\partial f(\bar{x})$  is not a singleton, the above property has to be satisfied just by *some* of its elements: this one element, call it  $s_1$ , suffices to rule out any feasible descent direction. Indeed,

$$f(\bar{x} + td) - f(\bar{x}) \geq t \langle s_1, d \rangle \quad \text{for all } t \geq 0;$$

if  $s_1$  satisfies the above angle property, these terms are nonnegative for all feasible  $d$ .



**Fig. 1.1.1.** The dual minimality condition

**Remark 1.1.3** If the problem were to maximize  $f$  over  $C$ , the (then local and not sufficient) optimality condition (ii) = (ii') would become

$$\begin{aligned} f'(\bar{x}, y - \bar{x}) &\leq 0 \quad \text{for all } y \in C, \\ f'(x, d) &\leq 0 \quad \text{for all } d \in T_C(\bar{x}). \end{aligned}$$

This would mean that *all* subgradients should make an acute angle with the feasible directions for  $C$  at  $\bar{x}$ , i.e. (iii) = (1.1.3) should at least be replaced by

$$\partial f(\bar{x}) \subset N_C(\bar{x}) \quad (1.1.4)$$

– which would still be insufficient (because local no longer implies global), but at least more accurate.

Thus, maximizing a convex function over a convex set is a totally different problem, even though the data still enjoy the same properties.  $\square$

For an illustration of Theorem 1.1.1, suppose that  $f$  happens to be differentiable at  $\bar{x}$ , and let

$$C := \text{co}\{v_1, \dots, v_m\}$$

be a compact convex polyhedron characterized as a convex hull. In this case, it is condition (ii) that is the most useful: indeed, it reads

$$\langle \nabla f(\bar{x}), y - \bar{x} \rangle \geq 0 \quad \text{for all } y \in C.$$

It is immediately seen that this is true if and only if

$$\langle \nabla f(\bar{x}), v_j - \bar{x} \rangle \geq 0 \quad \text{for } j = 1, \dots, m,$$

a set of conditions very easy to check.

**Example 1.1.4 (Affine Manifolds)** Let  $C = \{x_0\} + H$  be an affine manifold in  $\mathbb{R}^n$ ,  $H$  being a subspace. The tangent and normal cones at  $x$  to  $C$  are also the tangent and normal cone at  $x - x_0$  to  $H$ , namely  $H$  itself and its orthogonal  $H^\perp$  respectively. In this case, the primal minimality condition (ii') is:  $f'(\bar{x}, d) \geq 0$  for all  $d \in H$ ; the dual minimality condition is:

there is a subgradient of  $f$  that is orthogonal to  $H$ .

Depending on the analytic description of  $H$ , this condition may take several forms.

If  $H$  is characterized as a linear hull, say: for given  $x_0$  and  $e_1, \dots, e_m$ ,

$$C = \left\{ x_0 + \sum_{j=1}^m \xi_j e_j : \xi = (\xi_1, \dots, \xi_m) \in \mathbb{R}^m \right\},$$

then (iii) is:

there is  $s \in \partial f(\bar{x})$  such that  $\langle s, e_j \rangle = 0$  for  $j = 1, \dots, m$ .

On the other hand,  $H$  can be characterized as an intersection of hyperplanes:  $A$  being a linear operator from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ , and  $b \in \mathbb{R}^m$ ,

$$C = \{x \in \mathbb{R}^n : Ax = b\} \quad (\text{here } Ax_0 = b). \quad (1.1.5)$$

Then  $H^\perp$  is the subspace  $\text{Im } A^*$  and (iii) becomes:  $A^*\lambda \in \partial f(\bar{x})$  for some  $\lambda \in \mathbb{R}^m$ , or

there are  $s \in \partial f(\bar{x})$  and  $\lambda \in \mathbb{R}^m$  such that  $s + A^*\lambda = 0$ . (1.1.6)

Note in this last expression that, even if  $s$  is fixed in  $\partial f(\bar{x})$ , there are as many possible  $\lambda$ 's as elements in  $\text{Ker } A^*$ . □

**Example 1.1.5** As a follow-up of the previous example, suppose again that  $C$  is characterized by (1.1.5), with  $A$  surjective, and take a quadratic objective function:

$$f(x) = \frac{1}{2} \langle Qx, x \rangle + \langle c, x \rangle,$$

with  $Q : \mathbb{R}^n \rightarrow \mathbb{R}^n$  symmetric positive definite and  $c \in \mathbb{R}^n$ . Clearly enough,  $C$  is nonempty and the minimization problem has a unique solution  $\bar{x}$ ; also, (1.1.6) has a unique solution  $(s, \lambda)$ , with  $s = Q\bar{x} + c$ . Let us compute  $\bar{x}$  and  $\lambda$ : they solve

$$\begin{aligned} Q\bar{x} + c + A^*\lambda &= 0, \\ A\bar{x} &= b. \end{aligned}$$

Since  $Q$  is invertible, this is equivalent to

$$\begin{aligned}\bar{x} + Q^{-1}c + Q^{-1}A^*\lambda &= 0, \\ A Q^{-1}A^*\lambda + A Q^{-1}c + b &= 0.\end{aligned}$$

The linear operator  $AQ^{-1}A^* : \mathbb{R}^m \rightarrow \mathbb{R}^m$  is clearly symmetric and we claim that it is positive definite: indeed, with  $\langle \cdot, \cdot \rangle$  denoting the scalar product in  $\mathbb{R}^m$ ,

$$\langle A Q^{-1}A^*y, y \rangle = \langle Q^{-1}A^*y, A^*y \rangle$$

is always nonnegative ( $Q^{-1}$  is positive definite), and it is positive if  $y \neq 0$  ( $A^*$  is injective). We therefore obtain the explicit expression

$$\lambda = -B(AQ^{-1}c + b), \quad \bar{x} = Q^{-1}A^*B(AQ^{-1}c + b) - Q^{-1}c,$$

where we have set  $B := (AQ^{-1}A^*)^{-1}$ .  $\square$

**Example 1.1.6 (Nonnegativity Constraints)** Suppose that  $\langle \cdot, \cdot \rangle$  is the usual dot-product and that

$$C = \{x = (\xi^1, \dots, \xi^n) : \xi^i \geq 0 \text{ for } i = 1, \dots, n\}$$

is the nonnegative orthant in  $\mathbb{R}^n$ . The expression of the normal cone to this  $C$  at a given  $x$  was given in Examples III.3.2.2(b) and III.5.2.6(a). We find that  $\bar{x}$  minimizes  $f$  on  $C$  if and only if there exists  $s = (s^1, \dots, s^n) \in \partial f(\bar{x})$  such that

$$\text{for } i = 1, \dots, n, \quad s^i \geq 0 \text{ and } s^i = 0 \text{ if } \xi^i > 0.$$

A slightly more complicated example is when

$$C = \Delta_n = \left\{ (\xi^1, \dots, \xi^n) : \sum_{i=1}^n \xi^i = 1, \xi^i \geq 0 \text{ for } i = 1, \dots, n \right\}$$

is the unit simplex of  $\mathbb{R}^n$ ; its normal cones were given in Example III.5.2.6(c). We obtain:  $\bar{x}$  minimizes  $f$  on  $C$  if and only if there exist  $s = (s^1, \dots, s^n) \in \partial f(\bar{x})$  and  $\lambda \in \mathbb{R}$  such that

$$s^i \geq -\lambda \text{ if } \xi^i = 0 \quad \text{and} \quad s^i = -\lambda \text{ if } \xi^i > 0. \quad \square$$

**Example 1.1.7 (“Oblique” Projections)** Let  $\|\cdot\|$  be a norm on  $\mathbb{R}^n$  and consider the problem of projecting a given  $x$  onto  $C$ : to find  $\bar{x} \in C$  (not necessarily unique!) such that

$$\|\bar{x} - x\| = \min_{y \in C} \|y - x\|. \quad (1.1.7)$$

Denote by

$$B^* := \{s : \langle s, y \rangle \leq \|y\| \text{ for all } y \in \mathbb{R}^n\}$$

the unit ball of the dual norm (see §V3.2). We have seen in Example VI.3.1 that the subdifferential of  $\|\cdot\|$  at  $z$  is the optimal set in  $B^*$ :

$$\partial\|\cdot\|(z) = \{s \in B^* : \langle s, z \rangle = \|z\|\} = \operatorname{Argmax}_{s \in B^*} \langle s, z \rangle.$$

Setting  $z = y - x$  in this formula, we can apply Theorem 1.1.1(iii):  $\bar{x}$  is a projection of  $x$  onto  $C$ , in the sense of the norm  $\|\cdot\|$ , if and only if

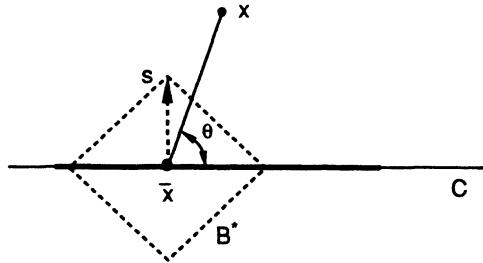
$$\exists s \in B^* \text{ such that } \langle s, x - \bar{x} \rangle = \|\bar{x} - x\| \text{ and } s \in N_C(\bar{x}).$$

Another way of expressing the same thing is that the problem

$$\max \{\langle s, x - \bar{x} \rangle : s \in B^*\} \quad (1.1.8)$$

has a solution in  $N_C(\bar{x})$ . Note that, if  $x \in C$ , then  $\bar{x} = x$  is itself a projection – unique because the minimal value in (1.1.7) is zero! In this case, the solution-set of (1.1.8) is the whole of  $B^*$ , which contains  $0 \in N_C(x)$ .

Figure 1.1.2 illustrates a situation in which  $C$  is the lower half-space in  $\mathbb{R}^2$  and  $\|\cdot\|$  is the  $\ell_\infty$ -norm. The projections of  $x$  are those  $\bar{x} = (\xi, 0)$  that see  $x$  under an angle  $\theta$  of at least  $\pi/4$ , so that the optimal  $s$  in (1.1.7) is the vertical  $(0, 1)$ . The projection-set is also  $B_\infty \cap C$ , where  $B_\infty$  is the  $\ell_\infty$ -ball around  $x$  having radius just equal to the  $\ell_\infty$ -distance from  $x$  to  $C$ .



**Fig. 1.1.2.** Solutions to a projection problem

If  $\|\cdot\|$  concides with the Euclidean norm  $\|\cdot\|$ , then  $B = B^*$  and

$$\partial\|\cdot\|(z) = \begin{cases} B & \text{if } z = 0, \\ \left\{ \frac{1}{\|z\|}z \right\} & \text{if } z \neq 0. \end{cases}$$

Using this value in (iii) shows that the projection onto  $C$  is the point  $p_C(x) \in C$  (known to be unique) such that  $x - p_C(x) \in N_C(x)$ . This can also be seen from (1.1.8), which has the unique solution  $(x - \bar{x})/\|x - \bar{x}\|$  (assuming  $x \notin C$ ). We thus come back to results seen in §III.3.1.  $\square$

## 1.2 Conceptual Exact Penalty

The constrained minimization problem (1.0.1) can be viewed as the unconstrained minimization of  $f + I_C$ , where  $I_C$  is the indicator function of  $C$ . The new objective becomes extended-valued, however; we will not study the subdifferentials of such functions until Chap. X; furthermore, minimizing such a function is not a computationally tractable task (cf. Chap. II). On the other hand,  $I_C$  can be viewed as an infinite penalty imposed to the points outside  $C$ . A relevant idea is therefore to approximate it by an *external penalty* function, i.e. a function  $p$  satisfying

$$p(x) = \begin{cases} 0 & \text{if } x \in C, \\ p(x) > 0 & \text{if } x \notin C. \end{cases} \quad (1.2.1)$$

An example of such a function is the distance-function  $d_C$ . When  $p$  is on hand, the original problem can be replaced by the unconstrained one

$$\inf \{f(x) + p(x) : x \in \mathbb{R}^n\}. \quad (1.2.2)$$

Now, a first natural question is: to what extent can we replace (1.0.1) by the simpler (1.2.2)? We start with elementary properties of external penalty functions, which are actually independent of any assumption on  $C$  and  $f$ .

**Lemma 1.2.1** *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $C \subset \mathbb{R}^n$ , and  $p : \mathbb{R}^n \rightarrow \mathbb{R}$  satisfy (1.2.1); call  $S$  and  $S_p$  the solution-sets (possibly empty) of (1.0.1) and (1.2.2) respectively.*

- (i) *Any  $x_p \in S_p$  which belongs to  $C$  is also a solution of (1.0.1).*
- (ii)  *$S_p \supset S$  whenever  $S_p \cap S \neq \emptyset$ .*
- (iii) *If  $S_p \cap S \neq \emptyset$  and if the penalty function  $q$  is such that  $q(x) > p(x)$  for all  $x \notin C$ , then  $S_q = S$ .*

PROOF. Let  $x_p$  solve (1.2.2): in particular, for all  $x \in C$ ,

$$f(x_p) + p(x_p) \leq f(x) + p(x) = f(x).$$

If  $x_p \in C$ , the first term is  $f(x_p)$ ; (i) is proved.

To prove (ii), take  $x_p \in S_p \cap S$  and let  $\bar{x} \in S$ ;  $x_p$  and  $\bar{x}$  are both in  $C$  and

$$f(\bar{x}) + p(\bar{x}) = f(\bar{x}) \leq f(x_p) = f(x_p) + p(x_p) \leq f(x) + p(x)$$

for all  $x \in \mathbb{R}^n$ . Hence  $\bar{x} \in S_p$ .

Finally, let  $p$  and  $q$  be as stated in (iii) and take  $x_p \in S_p \cap S$ ; it is easily seen that  $x_p \in S_q$ , hence  $S \subset S_q$  by virtue of (ii). Conversely, let  $x_q \in S_q$ ; if we can show  $x_q \in C$ , the proof will be finished thanks to (i). Indeed, we have

$$f(x_q) + q(x_q) \leq f(x_p) + q(x_p) = f(x_p) = f(x_p) + p(x_p); \quad (1.2.3)$$

and if  $x_q \notin C$ ,

$$f(x_q) + q(x_q) > f(x_q) + p(x_q) \geq f(x_p) + p(x_p),$$

which contradicts (1.2.3).  $\square$

Thus, feasibility is the only possibly missing property for solutions of the *penalized problem* (1.2.2) to solve the original problem (1.0.1); and to recover feasibility, the best is to increase the penalty function. To do so, the usual technique is to choose it of the form  $\pi p$ , with a fixed “basic” penalty function  $p$ , and a possibly increasing *penalty coefficient*  $\pi \geq 0$ .

Having chosen  $p$ , one solves

$$\inf \{f(x) + \pi p(x) : x \in \mathbb{R}^n\}. \quad (1.2.4)$$

It may happen that, no matter how  $\pi$  is chosen, (1.2.4) has no solution in  $C$  – or even no solution at all. By contrast, the favourable case is described by the following property:

**Definition 1.2.2 (Exact Penalty)** Let the constrained minimization problem (1.0.1) have a nonempty solution-set. A penalty function  $p$  satisfying (1.2.1) is said to have the *exact penalty property* if there is  $\pi \geq 0$  such that (1.2.4) has a solution belonging to  $C$ .

An equivalent definition (Lemma 1.2.1) is that the solution-sets of (1.0.1) and (1.2.4) coincide for  $\pi$  large enough.  $\square$

This property does hold for at least one basic penalty function, namely the distance-function:

**Theorem 1.2.3** Let  $C \subset \mathbb{R}^n$  be nonempty closed convex and  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be convex. Then the following statements are equivalent when  $\bar{x} \in C$ :

- (i)  $\bar{x}$  minimizes  $f$  over  $C$ ;
- (ii) there exists  $\pi > 0$  such that  $\bar{x}$  minimizes  $f + \pi d_C$  over  $\mathbb{R}^n$ .

PROOF. It is clear that (ii)  $\Rightarrow$  (i) since  $d_C = 0$  over  $C$ . Now, take  $r > 0$  and let  $\pi > 0$  be a Lipschitz constant of  $f$  over  $B(\bar{x}, r)$  (§IV.3.1); we claim that  $\bar{x}$  minimizes  $f + \pi d_C$ . Because the projection operator  $p_C$  over the convex set  $C$  is nonexpansive (§III.3.1),

$$\|\bar{x} - y\| \geq \|p_C(\bar{x}) - p_C(y)\| = \|\bar{x} - p_C(y)\|.$$

Thus, for  $y \in B(\bar{x}, r)$ ,  $p_C(y)$  is also in  $B(\bar{x}, r)$  and we can use the local Lipschitz property of  $f$ :

$$f(y) - f(p_C(y)) \geq -\pi \|y - p_C(y)\| = -\pi d_C(y) \quad (1.2.5)$$

and we deduce, if (i) holds:

$$f(y) + \pi d_C(y) \geq f(p_C(y)) \geq f(\bar{x}) = f(\bar{x}) + \pi d_C(\bar{x}).$$

We have thus proved that  $\bar{x}$  minimizes the convex function  $f + \pi d_C$  on  $B(\bar{x}, r)$ , hence on the whole of  $\mathbb{R}^n$ .  $\square$

The above proof uses direct arguments only; with some more refined results from convex analysis, it can be substantially shortened. Indeed,  $\bar{x}$  minimizes the convex function  $f + \pi d_C$  over  $\mathbb{R}^n$  if and only if (use the calculus rule VI.4.1.1 and Example VI.3.3)

$$0 \in \partial(f + \pi d_C)(\bar{x}) = \partial f(\bar{x}) + \pi[N_C(\bar{x}) \cap B(0, 1)] = \partial f(\bar{x}) + N_C(\bar{x}) \cap B(0, \pi).$$

To say that this holds for some  $\pi$  is really to say that

$$0 \in \partial f(\bar{x}) + N_C(\bar{x}),$$

i.e. the properties stated in Theorem 1.2.3(ii) and Theorem 1.1.1(iii) are equivalent when  $\bar{x} \in C$ . However, our proof is preferable, because it lends itself to generalizations:

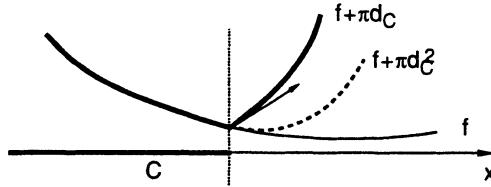
**Remark 1.2.4** It is interesting to extract the essential assumptions in Theorem 1.2.3:

- The only useful property from the projection operation  $p_C$  is its *local boundedness*, i.e. all the projections  $p_C(y)$  must be in some  $B(\bar{x}, R)$  if  $y \in B(\bar{x}, r)$ : then take for  $\pi$  in the proof a Lipschitz constant of  $f$  on  $B(\bar{x}, \max\{r, R\})$ .

- Apart from the local vs. global problem, convexity of  $f$  is of little use; what really matters for  $f$  is to be locally Lipschitzian.

Even under these weaker assumptions, the key inequality (1.2.5) still holds. The result is therefore valid under more general assumptions: for example, the projection can be made under some other metric; or  $d_C$  may be replaced by some other function behaving similarly; and convexity of  $f$  and  $C$  is secondary.

Observe that the exact penalty property is normally a concept attached to a particular  $f$ . The distance function, however, depends only on  $C$ ; it has therefore an “intrinsic exact penalty” property, which holds for arbitrary  $f$  (within a certain class, say  $f$  convex).  $\square$



**Fig. 1.2.1.** The property of exact penalty

The property of exact penalty is illustrated by Fig. 1.2.1. We see that it is essential for the penalizing function (here  $d_C$ ) to “break” the derivative of  $f$  when  $x$  crosses the boundary of  $C$ . In fact, another usual penalty technique replaces (1.0.1) by

$$\inf \left\{ f(x) + \frac{1}{2}\pi d_C^2(x) : x \in \mathbb{R}^n \right\}.$$

Here the property of exact penalty cannot hold in general, and  $\pi$  must really go to infinity: more precisely,

$$\partial \left( f + \frac{1}{2}\pi d_C^2 \right)(x) = \partial f(x) + \frac{1}{2}\pi \nabla d_C^2(x) = \partial f(x) \quad \text{for all } x \in C.$$

Hence, an  $x^* \in C$  minimizing the above penalized function should already satisfy  $0 \in \partial f(x^*)$ ; an uninteresting situation, in which (1.0.1) is an essentially unconstrained problem. Basically, the trouble is that the function  $d_C^2$  is smooth and its gradient is 0 for  $x \in C$ : when  $x$  leaves  $C$ ,  $\nabla d_C^2(x)$  is small and  $d_C^2(x)$  does not increase fast enough.

## 2 Minimality Conditions Involving Constraints Explicitly

Now, we suppose that the constraint-set of §1 has a *representation* via equalities and inequalities:  $C$  is the set of  $x \in \mathbb{R}^n$  such that

$$\langle a_i, x \rangle = b_i \text{ for } i = 1, \dots, m, \quad c_j(x) \leq 0 \text{ for } j = 1, \dots, p. \quad (2.0.1)$$

Here each  $(a_i, b_i) \in \mathbb{R}^n \times \mathbb{R}$ ,  $c_j : \mathbb{R}^n \rightarrow \mathbb{R}$  is a convex function; altogether, they form the data  $(a, b, c)$ .

The two groups of constraints in (2.0.1) really represent a classification “equalities vs. inequalities” – rather than “affine vs. general nonlinear”. In a way, it is by chance that the first group contains only affine functions: as far as equality constraints are

concerned to characterize a convex set, only affine functions are relevant; but notationally, we could just write the equalities as  $d_i(x) = 0$ , say. Note also that an equality could as well be written as a pair of affine inequalities. Similarly, some inequalities  $c_j$  may be affine, they still appear in the second group.

The following conventions will be useful:

- $m = 0$  means that the representation (2.0.1) has no equalities, while  $p = 0$  means that there are no inequalities.
- Since this last case has already been dealt with in Example 1.1.4, the present section will be essentially limited to two cases:  $[m = 0, p \geq 1]$  (only inequalities) and  $[m \geq 1, p \geq 1]$  (both types of constraints present).
- In either case, an expression like  $\sum_{i=1}^0$  means a summation on the empty set, whose result is by convention 0.

We find it convenient to equip  $\mathbb{R}^m$ , the space of equality-constraints values, with the standard dot-product: for  $(\lambda, b) \in \mathbb{R}^m \times \mathbb{R}^m$ ,

$$\lambda^\top b = \sum_{i=1}^m \lambda_i b_i. \quad (2.0.2)$$

Also,  $A : \mathbb{R}^n \rightarrow \mathbb{R}^m$  is the linear operator which, to  $x \in \mathbb{R}^n$ , associates the vector of coordinates  $\langle a_i, x \rangle$ ,  $i = 1, \dots, m$ . Thus we can write

$$Ax = b \quad \text{instead of} \quad [\langle a_i, x \rangle = b_i \text{ for } i = 1, \dots, m]$$

and

$$C = \{x \in \mathbb{R}^n : Ax = b, c_j(x) \leq 0 \text{ for } j = 1, \dots, p\}.$$

The adjoint  $A^*$  of  $A$  is then the operator which, to  $\lambda = (\lambda_1, \dots, \lambda_m) \in \mathbb{R}^m$ , associates the vector  $A^*\lambda = \sum_{i=1}^m \lambda_i a_i \in \mathbb{R}^n$ .

Thus, our basic convex minimization problem (1.0.1) is now written as

$$\begin{aligned} \min f(x) \quad & x \in \mathbb{R}^n, \\ \langle a_i, x \rangle = b_i \quad & \text{for } i = 1, \dots, m, \quad [\text{or } Ax = b \in \mathbb{R}^m], \\ c_j(x) \leq 0 \quad & \text{for } j = 1, \dots, p. \end{aligned} \quad (2.0.3)$$

Of course, the same set  $C$  can be represented by equalities and inequalities in many different ways. Just to give an example, there holds in terms of the distance-function  $d_C$ :

$$C = \{x \in \mathbb{R}^n : d_C(x) \leq 0\}, \quad (2.0.4)$$

or also

$$C = \left\{x \in \mathbb{R}^n : \frac{1}{2} d_C^2(x) \leq 0\right\}.$$

Thus, if  $d_C$  is known, we have already two possible representations in the form (2.0.1). As another example, which will be of fundamental importance below, consider the following summarization of the data in (2.0.1):

$$\begin{aligned} \mathbb{R}^n \ni x \mapsto \Gamma(x) := \\ \left( |\langle a_1, x \rangle - b_1|, \dots, |\langle a_m, x \rangle - b_m|; c_1^+(x), \dots, c_p^+(x) \right) \in \mathbb{R}^{m+p} \end{aligned} \quad (2.0.5)$$

(recall that  $t^+ := \max\{0, t\}$ ). It allows the description of  $C$  with the help of the unique vector-equation  $\Gamma(x) = 0$ . Taking an arbitrary norm  $\|\cdot\|$  in  $\mathbb{R}^{m+p}$ , this can then be expressed as

$$C = \{x \in \mathbb{R}^n : \gamma(x) \leq 0\}, \quad (2.0.6)$$

where  $\gamma(x) := \|\Gamma(x)\|$ .

We will see that the relevant object associated with the representation (2.0.6) is actually independent of the particular norm; we can for example choose

$$\gamma_\infty := \max \left\{ |\langle a_1, \cdot \rangle - b_1|, \dots, |\langle a_m, \cdot \rangle - b_m|; c_1^+, \dots, c_p^+ \right\}, \quad (2.0.7)$$

or

$$\gamma_1 := \sum_{i=1}^m |\langle a_i, \cdot \rangle - b_i| + \sum_{j=1}^p c_j^+. \quad (2.0.8)$$

In both cases  $\gamma$  is a (convex) “global constraint-function” characterizing  $C$ .

There are two ways (at least) of deriving minimality conditions in (2.0.3): one, which will be the subject of §2.1 – 2.3, is to use §1.1 after a characterization of  $T_C$  and  $N_C$  in terms of the data  $(a, b, c)$ ; the other, tackling the minimization problem (2.0.3) directly, will come in §2.4.

## 2.1 Expressing the Normal and Tangent Cones in Terms of the Constraint-Functions

When representing the convex set  $C$  by (2.0.6), it is desirable that  $\gamma$  be convex; for this, we choose a norm in  $\mathbb{R}^{m+p}$  satisfying: for all pairs  $(z, z') \in \mathbb{R}^{m+p} \times \mathbb{R}^{m+p}$ ,

$$\|z\| \leq \|z'\| \quad \text{whenever} \quad 0 \leq z^i \leq z'^i \text{ for all } i = 1, \dots, m+p \quad (2.1.1)$$

See §VI.4.3 for the convexity of the resulting function  $\gamma$ ; observe also that the  $\ell_p$ -norms,  $1 \leq p \leq \infty$ , satisfy this property. The subdifferential  $\partial\gamma(x)$  is then a convex compact set in  $\mathbb{R}^n$  which contains 0 if  $x \in C$ . Of course, this set depends on the norm  $\|\cdot\|$  chosen but its conical hull cone  $\mathbb{R}^+\partial\gamma(x) = \mathbb{R}^+\partial\gamma(x)$  does not:

**Lemma 2.1.1** *For  $i = 1, 2$ , let  $\|\cdot\|_i$  be two norms in  $\mathbb{R}^{m+p}$ , satisfying the monotonicity property (2.1.1), and let  $\gamma_i = \|\Gamma\|_i$  be the corresponding convex functions used in (2.0.5), (2.0.6). For any  $x \in C$ ,*

$$\mathbb{R}^+\partial\gamma_1(x) = \mathbb{R}^+\partial\gamma_2(x).$$

PROOF. The two norms are equivalent: there exist  $0 < \ell \leq L$  such that  $\ell\|\cdot\|_1 \leq \|\cdot\|_2 \leq L\|\cdot\|_1$ . As a result, for all  $x \in C$  (i.e.  $\gamma_i(x) = 0$ ),  $d \in \mathbb{R}^n$  and  $t > 0$ ,

$$\ell \frac{\gamma_1(x + td) - \gamma_1(x)}{t} \leq \frac{\gamma_2(x + td) - \gamma_2(x)}{t} \leq L \frac{\gamma_1(x + td) - \gamma_1(x)}{t}.$$

Let  $t \downarrow 0$  to obtain

$$\ell\gamma'_1(x, d) \leq \gamma'_2(x, d) \leq L\gamma'_1(x, d) \quad \text{for all } d \in \mathbb{R}^n.$$

According to the definition of a subdifferential, for example VI.1.1.4, this just means

$$\ell\partial\gamma_1(x) \subset \partial\gamma_2(x) \subset L\partial\gamma_1(x). \quad \square$$

It so happens that the conical hull considered in the above lemma is the useful object for our purpose. Its calculation turns out to be simplest if  $\|\cdot\|$  is taken as the  $\ell_1$ -norm; we therefore set

$$\mathbb{R}^n \ni x \mapsto \gamma(x) := \sum_{i=1}^m |\langle a_i, x \rangle - b_i| + \sum_{j=1}^p c_j^+(x),$$

but once again, any other norm would result in the same expression (2.1.3) below. We also recall the notation  $A^*\lambda$  for  $\sum_i \lambda_i a_i$ .

In what follows, we will denote by

$$J(x) := \{j = 1, \dots, p : c_j(x) = 0\} \quad (2.1.2)$$

the set of *active* inequality constraints at  $x \in C$  (or active set for short).

**Proposition 2.1.2** *For  $x \in C$ , the conical hull of  $\partial\gamma(x)$  is*

$$N'_{(a,b,c)}(x) := \left\{ A^*\lambda + \sum_{j \in J(x)} \mu_j s_j : \lambda \in \mathbb{R}^m, \mu_j \geq 0, s_j \in \partial c_j(x) \text{ for } j \in J(x) \right\}. \quad (2.1.3)$$

PROOF. Use the various relevant calculus rules in §VI.4 to obtain successively:

$$\begin{aligned} \partial(|\langle a_i, \cdot \rangle - b_i|)(x) &= [-1, +1]a_i \quad \text{for } i = 1, \dots, m; \\ \partial c_j^+(x) &= \begin{cases} [0, 1]\partial c_j(x) & \text{if } j \in J(x), \\ \{0\} & \text{if } j \notin J(x); \end{cases} \\ \partial\gamma(x) &= \sum_{i=1}^m [-1, +1]a_i + \sum_{j \in J(x)} [0, 1]\partial c_j(x), \end{aligned} \quad (2.1.4)$$

and (2.1.3) follows.  $\square$

The important point in (2.1.3) is that it involves only the data  $(a, b, c)$  of the problem: the cone  $N'_{(a,b,c)}(x)$  presents itself as a natural substitute for the normal cone  $N_C(x)$ . Now consider the polar of  $N'_{(a,b,c)}(x)$ , which is by definition

$$\begin{aligned} \{d \in \mathbb{R}^n : \langle s, d \rangle \leq 0 \text{ for all } s \in \mathbb{R}^+ \partial\gamma(x)\} &= \\ \{d \in \mathbb{R}^n : \langle s, d \rangle \leq 0 \text{ for all } s \in \partial\gamma(x)\} &= \\ \{d \in \mathbb{R}^n : \gamma'(x, d) \leq 0\} &=: T'_{(a,b,c)}(x). \end{aligned}$$

To alleviate notation, we will often write  $N'(x)$  and  $T'(x)$ , or even  $N'$  and  $T'$ , instead of  $N'_{(a,b,c)}(x)$  and  $T'_{(a,b,c)}(x)$ . Using for example (2.1.4), we can compute

$$\gamma'(x, d) = \sum_{i=1}^n |\langle a_i, d \rangle| + \sum_{j \in J(x)} c'_j(x, d),$$

so that, not unexpectedly,  $T'$  also has an expression in terms of the data of the problem only:

$$[N'(x)]^\circ = T'(x) = \{d \in \mathbb{R}^n : Ad = 0, c'_j(x, d) \leq 0 \text{ for } j \in J(x)\}. \quad (2.1.5)$$

Being a polar cone,  $T'$  is closed and convex; geometrically, it is obtained by linearizing – or sublinearizing – the constraints at  $x$ , a rather natural operation. When passing to the dual, the next natural operation is to take the polar of  $T'$ , but be aware that this does not give  $N'$ , simply because  $N'$  is not closed.

In §1 we used  $(T_C, N_C, d_C)$  only; here the corresponding triple is  $(T', N', \gamma)$ , attached to the data  $(a, b, c)$ . Our duty is now to make the connection between these two triples. The form (2.0.6) expresses  $C$  as a sublevel-set, so the problem of calculating its normal and tangent cones has already been addressed in §VI.1.3.

**Lemma 2.1.3** *For all  $x \in C$ , there holds*

$$T_C(x) \subset T'_{(a,b,c)}(x). \quad (2.1.6)$$

Furthermore,  $[T'_{(a,b,c)}(x)]^\circ = \text{cl } N'_{(a,b,c)}(x)$  and

$$N_C(x) \supset \text{cl } N'_{(a,b,c)}(x) \supset N'_{(a,b,c)}(x). \quad (2.1.7)$$

PROOF. (2.1.6) is just a rewriting of Lemma VI.1.3.2. Because  $T' = (N')^\circ$ , we have  $(T')^\circ = (N')^{\circ\circ} = \text{cl } N'$  and (2.1.7) is obtained by taking the polar of both sides in (2.1.6).  $\square$

We know from Theorem 1.1.1 that a solution  $\bar{x}$  of our minimization problem (2.0.3) is characterized by

$$f'(\bar{x}, d) \geq 0 \text{ for all } d \in T_C(\bar{x}), \quad \text{or} \quad 0 \in \partial f(\bar{x}) + N_C(\bar{x});$$

but we wish to use the data  $(a, b, c)$ . This amounts to writing

$$f'(\bar{x}, d) \geq 0 \text{ for all } d \in T'(\bar{x}), \quad \text{or} \quad 0 \in \partial f(\bar{x}) + N'(\bar{x})$$

(with an extra technical detail:  $N'$  is not the polar cone of  $T'$ ). We therefore need the property  $N'(\bar{x}) = N_C(\bar{x})$ , which may not hold in general. Nevertheless, Lemma 2.1.3 enables us to prove the following fundamental result:

**Theorem 2.1.4** *For  $\bar{x} \in C$ , consider the following statements:*

- (i)  $\bar{x}$  solves the constrained minimization problem (2.0.3);

- (ii)  $f'(\bar{x}, d) \geq 0$  for all  $d \in T'_{(a,b,c)}(\bar{x})$ ;
- (iii)  $0 \in \partial f(\bar{x}) + \text{cl } N'_{(a,b,c)}(\bar{x})$ ;
- (iv) there exist  $\lambda = (\lambda_1, \dots, \lambda_m) \in \mathbb{R}^m$  and  $\mu = (\mu_1, \dots, \mu_p) \in \mathbb{R}^p$  such that

$$0 \in \partial f(\bar{x}) + \sum_{i=1}^m \lambda_i a_i + \sum_{j=1}^p \mu_j \partial c_j(\bar{x}), \quad (2.1.8)$$

$$\mu_j \geq 0 \text{ and } \mu_j c_j(\bar{x}) = 0 \text{ for } j = 1, \dots, p. \quad (2.1.9)$$

Then we have the following relations: (iv)  $\Rightarrow$  (iii)  $\Leftrightarrow$  (ii)  $\Rightarrow$  (i). If the equality  $N' = N_C(\bar{x})$  holds, we have the full equivalence (i)  $\Leftrightarrow$  (ii)  $\Leftrightarrow$  (iii)  $\Leftrightarrow$  (iv).

PROOF. [(ii)  $\Leftrightarrow$  (iii)] Because  $T'$  and  $\text{cl } N'$  are mutually polar cones, this is the same as the equivalence between (ii') and (iii) in Theorem 1.1.1.

[(iv)  $\Rightarrow$  (iii)] Using the definition (2.1.3) of  $N'$ , (iv) means  $0 \in \partial f(\bar{x}) + N'$ , which itself implies (iii).

[(iii)  $\Rightarrow$  (i)] In view of (2.1.7), (iii) implies  $0 \in \partial f(\bar{x}) + N_C(\bar{x})$  which, according to Theorem 1.1.1, means that  $\bar{x}$  minimizes  $f$  over  $C$ .

Finally, the equality  $N' = N_C(\bar{x})$  implies in particular  $N' = \text{cl } N'$ , and also  $T' = T_C(\bar{x})$ ; the four statements become equivalent, just as in Theorem 1.1.1.  $\square$

The statements (i), (ii), (iii) in this result play the role of those in Theorem 1.1.1; as for (iv), it does nothing other than develop the expression of  $N'$ , thereby giving a computable way of checking the condition  $0 \in \partial f + N'$ . In Theorem 2.1.4, the difference between (ii) = (iii) and (iv) is slim: only the boundary of  $N'$  is involved (see Example 2.1.7 below, though). The real question is whether (i) implies (ii) = (iii)  $\simeq$  (iv): then, a computable necessary condition is obtained to eliminate a candidate  $\bar{x}$  which would not be optimal. If this implication does not hold, our computable condition (iv) [ $\simeq$  (iii) = (ii)] is only sufficient for optimality.

The equivalence between (i) and (ii) = (iii) is given by the property  $N'(\bar{x}) = N_C(\bar{x})$ , which yields the closedness of  $N'$  at the same time – hence (iv). This property thus appears as a cornerstone to derive conditions equivalent to minimality; it will motivate Sections 2.2, 2.3 by itself.

The existence of coefficients satisfying (2.1.8), (2.1.9) in Theorem 2.1.4 is called *Lagrange*, or *Karush-Kuhn-Tucker* (KKT) conditions; actually, Lagrange derived them in the case of equality constraints only, and for differentiable data. The corresponding coefficients  $(\lambda, \mu) \in \mathbb{R}^m \times (\mathbb{R}^+)^p$  are called the (Lagrange) *multipliers*.

**Remark 2.1.5** Call  $c := (c_1, \dots, c_p) \in \mathbb{R}^p$ . For feasible  $x$ , the vector  $-c(x)$  is in  $(\mathbb{R}^+)^p$  and there are different ways of expressing (2.1.9):

- one is  $\mu \in (\mathbb{R}^+)^p$  and  $\mu_j = 0$  whenever  $c_j(\bar{x}) < 0$  (but the converse is not true: one may have  $\mu_j$  and  $c_j(\bar{x})$  both zero);
- another one is  $\mu \in (\mathbb{R}^+)^p$  and  $\mu^\top c(\bar{x}) = 0$ .

This equivalence, together with the notation (2.1.2), allows the following abbreviation of (2.1.8) and (2.1.9):

$$0 \in \partial f(\bar{x}) + A^* \lambda + \sum_{j \in J(\bar{x})} \mu_j \partial c_j(\bar{x}), \quad \mu_j \geq 0 \text{ for } j \in J(\bar{x}). \quad (2.1.10)$$

The condition  $\mu^\top c = 0$  is called *transversality*, or *complementarity slackness*; when  $c_j(\bar{x}) = 0$  does imply  $\mu_j > 0$ , we say that *strict complementarity slackness* holds.  $\square$

We finish this subsection with some illustrations.

**Example 2.1.6** If  $c = d_C$ , the condition  $N' = N_C$  obviously holds, since  $\partial d_C(x) = N_C(x) \cap B(0, 1)$  for feasible  $x$  (see Example VI.3.3). By contrast, representing  $C$  with the single inequality  $\frac{1}{2}d_C^2(x) \leq 0$  results in  $N' = \{0\}$ , probably a gross underestimate of the true normal cone  $N_C$ . This remark has a general interest: replacing a representation

$$C = \{x \in \mathbb{R}^n : c(x) \leq 0\}$$

by

$$C = \{x \in \mathbb{R}^n : \frac{1}{2}(c^+)^2(x) \leq 0\}$$

kills the possibility of having  $N' = N_C$  on the boundary of  $C$ . See again the quadratic penalty mentioned at the end of §1.2.  $\square$

**Example 2.1.7 (Nonclosed  $N'$ )** Take the dot-product for  $\langle \cdot, \cdot \rangle$  in  $\mathbb{R}^2$  and

$$C = \{x = (\xi, \eta) : c(x) := \xi + \|\xi, \eta\| \leq 0\} = \mathbb{R}^- \times \{0\}.$$

At  $x = 0$ , straightforward calculations give

$$T'(0) = \{d \in \mathbb{R}^2 : c'(0, d) \leq 0\} = C = T_C(0).$$

Then, a function  $f$  (whatever it is) is minimized on  $C$  at  $x = 0$  if and only if (ii) holds in Theorem 2.1.4. Yet,  $\partial c(0) = \{(1, 0)\} + B(0, 1)$  and

$$N'(0) = \mathbb{R}^+ \partial c(0) = \{s = (\rho, \tau) : \rho > 0\} \cup \{(0, 0)\}$$

is not closed: it is only  $\text{cl } N'$  that coincides with  $N_C(0)$ . Indeed, take the objective function  $f(\xi, \eta) = \eta$  (which is constant on  $C$ , hence minimal at 0). Its gradient  $\nabla f(0) = (0, 1)$  is not in  $-N'$ ; (iv) does not hold in Theorem 2.1.4.

This phenomenon is entirely due to nonsmoothness: if the constraints  $c_j$  were smooth,  $N'$  would have finitely many generators, and as such would be closed (Farkas Lemma III.4.3.3).  $\square$

## 2.2 Constraint Qualification Conditions

In the previous section, we have seen that the relevant minimality condition in terms of the data  $(a, b, c)$  was the KKT conditions (2.1.8), (2.1.9), which needed the property

$$\boxed{N'_{(a,b,c)}(x) = N_C(x).} \quad (2.2.1)$$

This property will be called the *basic constraint qualification* condition (BCQ). It is of fundamental use in optimization theory and deserves additional comments.

- As shown in Example 2.1.6, it is a property to be enjoyed not by  $C$  itself, but by the constraints defining it: BCQ depends on the *representation* of  $C$ ; and also, it depends on the particular  $x \in C$ . Take for example  $n = 1$ ,  $C = [0, 1]$  defined by a single inequality constraint with

$$c(x) = \begin{cases} \max\{0, -1 + x\} & \text{if } x \geq 0 \\ \frac{1}{2}x^2 & \text{if } x \leq 0; \end{cases}$$

BCQ is satisfied at  $x = 1$  but not at  $x = 0$ .

- In the representation (2.0.1) of  $C$ ,  $N'$  remains unchanged if we change  $c_j$  to  $c_j^+$ , and/or to  $t_j c_j$  ( $t_j > 0$ ) and/or  $(a_i, b_i)$  to  $(-a_i, -b_i)$ , and/or  $\langle a_i, x \rangle - b_i = 0$  to the pair  $\langle a_i, x \rangle - b_i \leq 0$ ,  $\langle a_i, x \rangle - b_i \geq 0$ . At least, BCQ enjoys some coherence, since these changes do not affect  $C$ .
- The basic constraint qualification does not depend on  $f$ . When it holds at some  $\bar{x}$ , it therefore allows the derivation of the KKT conditions (2.1.8), (2.1.9) for any objective function that is minimized at the given  $\bar{x}$ . The following result shows that the converse is also true: in some sense, BCQ is a “minimal” condition.

**Proposition 2.2.1** *Let  $C$  have the representation (2.0.1). For any  $\bar{x} \in C$ , the following statements are equivalent:*

- For any convex function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  minimized on  $C$  at  $\bar{x}$ , there exist  $\lambda \in \mathbb{R}^m$  and  $\mu \in \mathbb{R}^p$  such that the KKT conditions (2.1.8), (2.1.9) hold at  $\bar{x}$ .*
- The basic constraint qualification condition  $N'_{(a,b,c)}(\bar{x}) = N_C(\bar{x})$  holds.*

PROOF. It suffices to prove that (i) implies (ii), i.e.  $N_C(\bar{x}) \subset N'(\bar{x})$ . Let  $s \in N_C(\bar{x})$  and consider the affine function  $x \mapsto f(x) = \langle -s, x - \bar{x} \rangle$ . By definition of the normal cone,  $f$  is minimized on  $C$  at  $\bar{x}$ ; then (i) applies, which means that  $0 \in -s + N'(\bar{x})$ .

□

We mention some situations where BCQ holds at every  $x \in C$ .

- When  $C$  is represented by  $d_C(x) \leq 0$ : Example 2.1.6 tells us that we recover Theorem 1.1.1.
- When  $m = 0$  (inequalities only) and there exists  $x_0$  such that  $c_j(x_0) < 0$  for  $j = 1, \dots, p$ : then, using Theorem VI.1.3.5 and the calculus rule VI.4.3.2 directly gives (for  $J(x) \neq \emptyset$ )

$$\begin{aligned} N_C(x) &= \mathbb{R}^+ \partial(\max_j c_j)(x) = \text{cone}\{\cup \partial c_j(x) : J \in J(x)\} = \\ &\left\{ \sum_{j \in J(x)} \mu_j s_j : \mu_j \geq 0, s_j \in \partial c_j(x) \text{ for } j \in J(x) \right\} = N'(x). \end{aligned}$$

- When  $C$  is represented by *affine* equalities and inequalities. This is indeed an important result:

**Proposition 2.2.2** *Let the constraint-set  $C$  have the representation*

$$C = \{x \in \mathbb{R}^n : Ax = b, \langle s_j, x \rangle - r_j \leq 0 \text{ for } j = 1, \dots, p\}.$$

*If  $J(x)$  denotes again the active set (2.1.2) at  $x \in C$ , we have*

$$N_C(x) = \text{Im } A^* + \text{cone}\{s_j : j \in J(x)\} = N'(x)$$

*and BCQ holds at every  $x \in C$ .*

PROOF. Use Example III.5.2.6(b) to obtain the above expression for the normal cone.  $\square$

The next problem will be to check the basic constraint qualification condition easily, in terms of the data  $(a, b, c)$ . A natural question is therefore: can BCQ be replaced by *more practical* (and possibly more restrictive) conditions? One such is the condition mentioned above in the context of inequalities, whose importance will be clear later. To state it, the affine constraints in (2.0.1) must be particularized:

$$J_a := \{j = 1, \dots, p : c_j \text{ is an affine function}\}$$

will denote the set (possibly empty) of affine inequality constraints.

**Definition 2.2.3** We say that the constraint-set (2.0.1) satisfies the *weak Slater assumption* (WSA for short) if there is a point at which all the non-affine constraints are strictly satisfied, i.e.:

$$\exists x_0 \in C \text{ such that } \begin{cases} Ax_0 = b, \\ c_j(x_0) \leq 0 \quad \text{for } j \in J_a, \\ c_j(x_0) < 0 \quad \text{for } j \notin J_a. \end{cases} \quad (2.2.2) \quad \square$$

**Proposition 2.2.4** An equivalent formulation of (2.2.2) is:

$$\forall x \in C, \exists d \in \mathbb{R}^n \text{ such that } \begin{cases} Ad = 0, \\ c'_j(x, d) \leq 0 \quad \text{for } j \in J(x) \cap J_a, \\ c'_j(x, d) < 0 \quad \text{for } j \in J(x) \setminus J_a. \end{cases} \quad (2.2.3)$$

PROOF.  $[(2.2.2) \Rightarrow (2.2.3)]$  Consider  $x \in C$  with  $J(x) \neq \emptyset$  (otherwise there is nothing to prove) and take  $d = x_0 - x$  ( $\neq 0$ ). Then  $Ad = Ax - Ax_0 = 0$ ; the inequality  $c'_j(x, d) \leq c_j(x_0) - c_j(x) = c_j(x_0)$ , true for all  $j \in J(x)$ , does the rest.

$[(2.2.3) \Rightarrow (2.2.2)]$  Consider  $x \in C$  with  $J(x) \neq \emptyset$  (if no such  $x$  exists, (2.2.2) just expresses  $C \neq \emptyset$ ) and compute  $c_j(x + td)$  for small  $t > 0$ . Since

$$c'_j(x, d) = \inf_{t>0} \frac{c_j(x + td)}{t} \quad \text{for } j \in J(x) \setminus J_a$$

and

$$c'_j(x, d) = \frac{c_j(x + td)}{t} \quad \text{for } j \in J(x) \cap J_a,$$

there exists  $t_0 > 0$  such that  $x_0 = x + t_0 d$  satisfies (2.2.2).  $\square$

It is interesting to note that (2.2.2) holds as soon as the conditions in (2.2.3) are satisfied by *some*  $x \in C$ . In such a case, these conditions are therefore satisfied for *all*  $x \in C$ ; this “propagation effect”, typical of convex constraint-sets, was already seen in Remark VI.1.3.6.

Thus, we have partitioned the inequality constraints into two sets:

- (i) the non-affine constraints  $c_j(x) \leq 0$  for all  $j \in \{1, \dots, p\} \setminus J_a$ , which could be summarized in a single inequality constraint

$$c_0(x) := \max \{c_j(x) : j \notin J_a\} \leq 0; \quad (2.2.4)$$

- (ii) the affine inequality constraints, to which could actually be added the “two-sided inequality versions” of the affine equality constraints:

$$\langle a_i, x \rangle - b_i \leq 0 \text{ and } \langle -a_i, x \rangle + b_i \leq 0 \quad \text{for } i = 1, \dots, m.$$

It is worth noting that the weak Slater assumption (2.2.2) remains unchanged under these transformations. So, *in fine*, the constraints describing  $C$  in (2.0.1) could be replaced by

- (i) a single (non-affine) inequality constraint  $c_0(x) \leq 0$ , and
- (ii) affine inequalities, say  $\langle s_k, x \rangle - r_k \leq 0$  for  $k = 1, \dots, q$ .

With these new notations, WSA is formulated as

$$\exists x_0 \in C \text{ such that } \begin{cases} \langle s_k, x_0 \rangle - r_k \leq 0 & \text{for } k = 1, \dots, q, \\ \text{and } c_0(x_0) < 0. \end{cases} \quad (2.2.5)$$

**Theorem 2.2.5** *Let the weak Slater assumption hold. At any  $\bar{x} \in C$ , the KKT conditions (2.1.8), (2.1.9) are (sufficient and) necessary for  $\bar{x}$  to minimize a convex function  $f$  on  $C$ .*

PROOF. Let  $\bar{x}$  minimize  $f$  on  $C$ , we have to show that the KKT conditions hold. Using the notation (2.2.5), consider the auxiliary function

$$\mathbb{R}^n \ni x \mapsto F(x) := \max \{f(x) - f(\bar{x}), c_0(x)\},$$

the closed convex polyhedron

$$P := \{x \in \mathbb{R}^n : \langle s_k, x \rangle - r_k \leq 0 \text{ for } k = 1, \dots, q\},$$

and the set of affine constraints

$$K(\bar{x}) := \{k = 1, \dots, q : \langle s_k, \bar{x} \rangle - r_k = 0\}.$$

Clearly enough,  $F(x) \geq F(\bar{x}) = 0$  for all  $x \in P$ . By virtue of Proposition 2.2.2, there exist nonnegative multipliers  $\mu_k$ ,  $k \in K(\bar{x})$ , such that

$$0 \in \partial F(\bar{x}) + \sum_{k \in K(\bar{x})} \mu_k s_k. \quad (2.2.6)$$

Let us compute  $\partial F(\bar{x})$ : if  $c_0(\bar{x}) < 0$ , then

$$\partial F(\bar{x}) = \partial[f - f(\bar{x})](\bar{x}) = \partial f(\bar{x});$$

if  $c_0(\bar{x}) = 0$ , use the calculus rule VI.4.3.2 to obtain

$$\partial F(\bar{x}) = \text{co}[\partial f(\bar{x}) \cup \partial c_0(\bar{x})].$$

In both cases, we deduce from (2.2.6) the existence of  $\alpha \in [0, 1]$  such that

$$0 \in \alpha \partial f(\bar{x}) + (1 - \alpha) \partial c_0(\bar{x}) + \sum_{k \in K(\bar{x})} \mu_k s_k. \quad (2.2.7)$$

We prove  $\alpha > 0$  (which is automatically true if  $c_0(\bar{x}) < 0$ ). Indeed, if  $\alpha$  were 0, we would have from (2.2.7)

$$0 \in \partial c_0(\bar{x}) + \sum_{k \in K(\bar{x})} \mu_k s_k.$$

This would imply that  $\bar{x}$  minimize  $c_0$  on  $P$  (Theorem 2.1.4). Because we are in the case  $c_0(\bar{x}) = 0$ , this would contradict (2.2.5).

In summary, dividing (2.2.7) by  $\alpha > 0$  if necessary, we have exhibited nonnegative multipliers  $v_0$  and  $v_k$ ,  $k \in K(\bar{x})$ , such that

$$0 \in \partial f(\bar{x}) + v_0 \partial c_0(\bar{x}) + \sum_{k \in K(\bar{x})} v_k s_k.$$

Finally, referring back to the original data  $(a, b, c)$  of our problem (2.0.3), we just remark that

- any  $s \in \partial c_0(\bar{x})$  can be written as a convex combination of elements in  $\partial c_j(\bar{x})$ ,  $j \in J(\bar{x}) \setminus J_a$ ;
- pairs of nonnegative multipliers, say  $(v_k, v_{k'})$ , representing an original affine equality, say the  $i^{\text{th}}$ , can be condensed in one unsigned multiplier  $\lambda_i := v_k - v_{k'}$ .  $\square$

Thus, WSA is a practical property ensuring the existence of Lagrange multipliers *for any pair*  $(\bar{x}, f)$  satisfying

$$f \text{ is convex from } \mathbb{R}^n \text{ to } \mathbb{R}, \text{ and } \bar{x} \in C \text{ minimizes } f \text{ on } C. \quad (2.2.8)$$

Remembering Proposition 2.2.1, we see that this implies the BCQ condition (2.2.1).

### 2.3 The Strong Slater Assumption

For given optimal  $\bar{x} \in C$ , denote by  $M(\bar{x})$  the set of multipliers. So far, we have concentrated our attention on conditions to guarantee *nonemptiness* of  $M(\bar{x})$ ; namely BCQ of (2.2.1), and WSA of (2.2.2). Now,  $M(\bar{x})$  is a *closed* set in  $\mathbb{R}^{m+p}$ , as can be checked from its definition; furthermore, it is *convex*: to see this, look again at the KKT conditions and remember that, because each  $\partial c_j(\bar{x})$  is convex,

$$\alpha \mu_j \partial c_j(\bar{x}) + (1 - \alpha) \mu'_j \partial c_j(\bar{x}) = [\alpha \mu_j + (1 - \alpha) \mu'_j] \partial c_j(\bar{x}) \quad \text{for } \alpha \in [0, 1].$$

To say more about  $M(\bar{x})$ , we need to restrict our qualification conditions; the following practical strengthening of WSA will imply among others that  $M(\bar{x})$  is also *bounded*.

**Definition 2.3.1** We say that the constraint-set (2.0.1) satisfies the *strong Slater assumption* (SSA for short) if:

the vectors $a_i$ , $i = 1, \dots, m$ , are linearly independent, $\exists x_0$ such that $Ax_0 = b$ and $c_j(x_0) < 0$ for $j = 1, \dots, p$ .	(i) (ii) $\square$
--	-----------------------

(2.3.1)

Thus the affine inequality constraints are no longer particularized. As a result, SSA would be killed if equalities were split into pairs of inequalities! The (i)-part, stating that  $A$  is surjective, is not too restrictive: it expresses the fact that the system  $Ax = b$  is not redundant. Needless to say,  $M(\bar{x})$  would certainly be unbounded otherwise: for  $(\lambda, \mu) \in M(\bar{x})$ , the whole affine manifold  $(\lambda + \text{Ker } A^*, \mu)$  would be in  $M(\bar{x})$ .

Just as in Proposition 2.2.4, the (ii)-part of SSA can be replaced by

$$\left. \begin{array}{l} \text{For all } x \in C, \text{ there exists } d \in \mathbb{R}^n \text{ such that} \\ Ad = 0 \quad \text{and} \quad c'_j(x, d) < 0 \text{ for all } j \in J(x); \end{array} \right\} \text{(ii') } \quad (2.3.1)$$

and here again, the required relations hold for all  $x \in C$  as soon as they hold at some  $x \in C$ .

As before, SSA is exclusively concerned with the description (2.0.1) of  $C$ . It will imply that  $M(\bar{x})$  is nonempty compact and convex for all pairs  $(\bar{x}, f)$  satisfying (2.2.8). Furthermore, the condition is necessary: one cannot have a pair  $(\bar{x}, f)$  satisfying (2.2.8) with  $M(\bar{x})$  nonempty compact and convex if SSA does not hold. This is summarized in the next statement:

**Theorem 2.3.2** *Consider  $\bar{x}$  minimizing  $f$  over the constraint-set  $C$  described in (2.0.1). A necessary and sufficient condition for the set of multipliers  $M(\bar{x})$  to be nonempty compact and convex is the strong Slater assumption (2.3.1).*

PROOF. [Sufficiency] Let SSA hold. We already know that  $M(\bar{x})$  is nonempty, closed and convex, we have to prove that it is bounded. For this, (2.3.1)(ii') allows us to take  $d \in \mathbb{R}^n$  such that

$$Ad = 0 \quad \text{and} \quad c'_j(\bar{x}, d) \leq -\varepsilon < 0 \text{ for } j \in J(\bar{x}).$$

Compute at this  $d$  the support function of the right-hand side in (2.1.8) – a non-negative number. Using various results from Chap. VI and knowing that  $Ad = 0$ , it has the value

$$f'(\bar{x}, d) + \sum_{j \in J(\bar{x})} \mu_j c'_j(\bar{x}, d) \geq 0.$$

Because each  $\mu_j$  is nonnegative, we obtain the bound for  $\mu$

$$\sum_{j=1}^p |\mu_j| = \sum_{j \in J(\bar{x})} \mu_j \leq \frac{f'(\bar{x}, d)}{\varepsilon}. \quad (2.3.2)$$

Let us now show the boundedness of the  $\lambda$ -contribution in  $M(\bar{x})$ . Consider the subspace  $E := \text{lin}(a_1, \dots, a_m)$  generated by the rows  $a_i$  of  $A$ . We write the KKT conditions as

$$\sum_{i=1}^m \lambda_i a_i \in -\partial f(\bar{x}) - \sum_{j \in J(\bar{x})} \mu_j \partial c_j(\bar{x}).$$

Because  $\partial f(\bar{x})$ , the  $\partial c_j(\bar{x})$ 's and the  $\mu_j$ 's are bounded, the right-hand side is a bounded set of the finite-dimensional space  $E$ , in which  $\{a_1, \dots, a_m\}$  is a basis by assumption. This implies that the corresponding coordinates  $\lambda_i$  are bounded.

[*Necessity*] Suppose that SSA does not hold: we have to prove that  $M(\bar{x})$  is unbounded if nonempty.

In case (2.3.1)(i) does not hold, we have already observed after Definition 2.3.1 that  $M(\bar{x})$  is either empty or unbounded. So suppose it is (2.3.1)(ii) which does not hold: for all  $x$  satisfying  $Ax = b$ , we have

$$c_0(x) := \max \{c_j(x) : j = 1, \dots, p\} \geq 0 \quad [= c_0(\bar{x})].$$

This implies that  $\bar{x}$  minimizes  $c_0$  over the affine manifold  $Ax = b$ . From Example 1.1.4 and the calculus rule VI.4.3.2, there exist  $\lambda' \in \mathbb{R}^m$  and convex multipliers  $\mu'_j$ ,  $j \in J(\bar{x})$ , such that

$$0 \in \sum_{i=1}^m \lambda'_i a_i + \sum_{j \in J(\bar{x})} \mu'_j \partial c_j(\bar{x}).$$

Thus, if  $(\lambda, \mu) \in M(\bar{x})$ , any element of the form  $(\lambda + t\lambda', \mu + t\mu')$  is again in  $M(\bar{x})$  for  $t \geq 0$ . Since  $\mu' \neq 0$ , this implies that  $M(\bar{x})$  is unbounded.  $\square$

**Remark 2.3.3** An effective bound for the  $\mu$ -part of the KKT multipliers can be derived: to obtain (2.3.2), we can take  $d = x_0 - \bar{x}$ , in which case

$$f'(\bar{x}, d) \leq f(x_0) - f(\bar{x}),$$

$$c'_j(\bar{x}, d) \leq c_j(x_0) \quad \text{for all } j \in J(\bar{x}),$$

hence

$$\sum_{j=1}^p \mu_j \leq \frac{f(x_0) - \bar{f}}{\min\{-c_j(x_0) : j = 1, \dots, p\}}. \quad \square$$

**Remark 2.3.4** It is interesting to note that, under SSA, we have

$$\text{ri } C = \{x \in \mathbb{R}^n : Ax = b \text{ and } c_j(x) < 0 \text{ for } j = 1, \dots, m\}$$

$$\text{rbd } C = \{x \in \mathbb{R}^n : Ax = b \text{ and } c_j(x) = 0 \text{ for some } j\}.$$

In the above expressions, the word ‘‘relative’’ can be dropped if there are no equality constraints, cf. Proposition VI.1.3.3.  $\square$

**Example 2.3.5** Consider the problem of finding a steepest-descent direction, which was the subject of §II.2.1. More precisely, let  $Q$  be a symmetric positive definite operator and take  $\|d\|^2 := \langle Qd, d \rangle$  for the normalization constraint in (II.2.1.3). With Remark II.2.1.4 in mind, we choose a normalization factor  $\kappa > 0$  and we want to find  $d$  solving

$$\min \{\langle s, d \rangle : \frac{1}{2} \langle Qd, d \rangle = \frac{1}{2}\kappa\} \tag{2.3.3}$$

(here  $s = s(x) \neq 0$  is the gradient of  $f$  at the given current iterate  $x$ ).

Consider the following relaxation of (2.3.3):

$$\min \{\langle s, d \rangle : \frac{1}{2} \langle Qd, d \rangle \leq \frac{1}{2}\kappa\}. \tag{2.3.4}$$

Now we have a convex minimization problem, obviously satisfying SSA – take  $d_0 = 0$ . According to Theorem 2.2.5,  $\bar{d}$  solves (2.3.4) if and only if there is  $\mu$  such that

$$s + \mu Q\bar{d} = 0, \quad \mu \geq 0, \quad \text{and} \quad \mu (\langle Q\bar{d}, \bar{d} \rangle - \kappa) = 0.$$

Because  $s \neq 0$ , this  $\mu$  cannot be zero:  $\langle Q\bar{d}, \bar{d} \rangle = \kappa$  and we can write

$$\bar{d} = -\frac{1}{\mu} Q^{-1}s \quad \text{and} \quad \langle Q\bar{d}, \bar{d} \rangle = \frac{1}{\mu^2} \langle s, Q^{-1}s \rangle = \kappa.$$

The last equation gives  $\mu$  as a function of  $\kappa$  (remember  $\mu > 0$ ), and  $\bar{d}$  is then obtained from the first equation. This  $\bar{d}$  solves (2.3.4) and is a posteriori feasible in (2.3.3); hence it solves (2.3.3) as well.

Observe in passing the confirmation of Remark II.2.1.4: as a direction, the optimal  $\bar{d}$  does not depend on  $\kappa > 0$ ; changing  $\kappa$  amounts to changing  $\mu$ , and just multiplies  $\bar{d}$  by some positive factor.  $\square$

## 2.4 Tackling the Minimization Problem with its Data Directly

So far, we have studied *sufficient* conditions only, namely the KKT conditions of Theorem 2.1.4; they became *necessary* under some qualification condition: BCQ of (2.2.1), or a practical property guaranteeing it, say some Slater assumption. All these qualification conditions involved the data  $(a, b, c)$  defining  $C$  in (2.0.1), but not  $f$ .

On the other hand, all the data  $(f, a, b, c)$  appearing in the minimization problem (2.0.3) can be collected into a set of minimality conditions, which are always *necessary*, without any assumption. The price to pay for this generality is that they are usually not too informative. Our basic tool for this will be the function (seen already in the proof of Theorem 2.2.5):

$$\mathbb{R}^n \ni x \mapsto F(x) := \max \{ f(x) - \bar{f}; c_1(x), \dots, c_p(x) \}. \quad (2.4.1)$$

**Proposition 2.4.1** *Let  $\bar{f}$  in (2.4.1) be the optimal value of (2.0.3). Then the problem*

$$\inf \{F(x) : Ax = b\}$$

*has optimal value 0 and the same solution-set as (2.0.3).*

**PROOF.** Straightforward: by definition,  $F(x) \geq 0$  for all  $x$  satisfying  $Ax = b$ , and to say that such an  $x$  satisfies  $F(x) = 0$  is to say that it solves (2.0.3).  $\square$

Thus, it suffices to write the minimality conditions of our new problem, from which all nonlinear constraints have been removed:

**Theorem 2.4.2** *If  $\bar{x}$  solves (2.0.3), there exist  $\lambda = (\lambda_1, \dots, \lambda_m) \in \mathbb{R}^m$  and  $(\mu_0, \mu) = (\mu_0, \mu_1, \dots, \mu_p) \in \mathbb{R} \times \mathbb{R}^p$ , with  $\mu_0$  and  $\mu$  not both zero, such that*

$$0 \in \mu_0 \partial f(\bar{x}) + \sum_{i=1}^m \lambda_i a_i + \sum_{j=1}^p \mu_j \partial c_j(\bar{x}), \quad (2.4.2)$$

$$\mu_j \geq 0 \text{ for } j = 0, 1, \dots, p \quad \text{and} \quad \mu_j c_j(\bar{x}) = 0 \text{ for } j = 1, \dots, p. \quad (2.4.3)$$

PROOF. We know that  $\bar{x}$  minimizes  $F$  on the affine manifold of equation  $Ax = b$ . Use Example 1.1.4 and the calculus rule VI.4.3.2 to compute  $\partial F(\bar{x})$  and obtain the required result, in which

$$\mu_0 + \sum_{j \in J(\bar{x})} \mu_j = 1.$$

□

The set of necessary conditions for minimality introduced by this result is called *John's conditions*; we will call the associated multipliers  $(\lambda, \mu_0, \mu) \in \mathbb{R}^{m+1+p}$  the *positively homogeneous* (or John's) multipliers; and we will call  $M_0(\bar{x})$  the (nonempty) set of such multipliers. Just as  $M(\bar{x})$ , it is a convex set; and it is obviously also a cone, which explains our wording “positively homogeneous”. On the other hand,  $M_0(\bar{x})$  is not closed because 0 has been excluded from it; but  $\{0\} \cup M_0(\bar{x})$  is indeed closed: proceed as with  $M(\bar{x})$ .

**Remark 2.4.3** Let  $x \in C$ ; then  $\partial F(x)$  is *not* the convex hull of  $\partial f(x)$  and of the  $\partial c_j(x)$ 's for active indices  $j$ : for this, we should have  $f(x) = \bar{f}$ . As a result, the minimality conditions in Theorem 2.4.2 do *not* state that  $\bar{x}$  minimizes  $F$  on the affine manifold  $Ax = b$ . These conditions are *not* sufficient: if, for example, there is some index  $j_0$  such that  $c_{j_0} \equiv 0$  on  $C$ , then any point in  $C$  satisfies John's conditions: set each  $\lambda_i$  and  $\mu_j$  to 0 except  $\mu_{j_0} = 1$ .

The trick is that the unknown  $\bar{f}$  does not appear in (2.4.2), (2.4.3), even though this value has its importance. Naturally, John's conditions become useful in two cases:

- When a posteriori  $f(x) = \bar{f}$ : for example,  $\bar{f}$  was known beforehand. Then the minimization of  $F$  is easy, see Example 1.1.4. Note, however, that (2.0.3) is no longer a constrained minimization problem, but rather a system of equations and inequations, namely

$$Ax = b, \quad f(x) \leq \bar{f}, \quad c_j(x) \leq 0 \text{ for } j = 1, \dots, p.$$

- When a posteriori a point has been found satisfying John's conditions with  $\mu_0 > 0$ : by positive homogeneity, we can take  $\mu_0 = 1$ . In other words,

$$M(x) = \{(\lambda, \mu) : (\lambda, 1, \mu) \in M_0(x)\}.$$

Here, we have gained nothing with respect to the standard KKT conditions. □

**Example 2.4.4** Take the constraint-set of Example 2.1.7:

$$C = \{x = (\xi, \eta) \in \mathbb{R}^2 : c(x) := \xi + \|(\xi, \eta)\| \leq 0\} = \mathbb{R}^- \times \{0\}$$

with the objective function

$$f_\alpha(\xi, \eta) = \eta - \alpha\xi \quad (\alpha \geq 0 \text{ being a parameter}).$$

In all cases,  $\bar{x} = 0$  minimizes  $f_\alpha$  over  $C$ .

If  $\alpha > 0$ , there are positively homogeneous multipliers with  $\mu_0 > 0$ , hence Lagrange multipliers: in fact,  $M(0) = [1/2(\alpha + 1/\alpha), +\infty[$ . If  $\alpha = 0$ ,  $M_0(0) = \{0\} \times \mathbb{R}_+^+$ . Naturally, and as predicted by Theorem 2.2.5, no Slater assumption holds in this example. □

Consider the minimization of  $f$  on an affine manifold of equation  $Ax = b$ , as in Example 1.1.4. There *are* Lagrange multipliers (WSA automatically holds); but there may exist “exotic” positively homogeneous multipliers, having  $\mu_0 = 0$ : those of the form  $(\lambda, 0)$ , with  $\lambda \in (\text{Ker } A^*) \setminus \{0\}$  (which implies  $A^*$  non-injective, and therefore precludes SSA). This explains the following result: generally speaking, to guarantee that *all* positively homogeneous multipliers have  $\mu_0 > 0$  precisely amounts to SSA.

**Theorem 2.4.5** *Let a pair  $(\bar{x}, f)$  satisfy (2.2.8). The strong Slater assumption (2.3.1) is equivalent to the property*

$$\mu_0 > 0 \quad \text{for all } (\lambda, \mu_0, \mu) \in M_0(\bar{x}). \quad (2.4.4)$$

PROOF. Using Theorem 2.3.2, we replace SSA by nonemptiness and boundedness of  $M(\bar{x})$ . Consider the sets (Fig. 2.4.1 is helpful)

$$H := \mathbb{R}^m \times \{1\} \times \mathbb{R}^p$$

and

$$K := \{0\} \cup M_0(\bar{x}) \subset \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}^p.$$

As observed already in Remark 2.4.3, their intersection

$$H \cap K = \{(\lambda, 1, \mu) : (\lambda, \mu) \in M(\bar{x})\}$$

is a shift of  $M(\bar{x})$ , and it is clear enough that  $M(\bar{x})$  is nonempty and bounded if and only if  $H \cap K$  is nonempty and bounded. From Proposition III.2.2.3, the latter holds if and only if the asymptotic cone  $(H \cap K)_\infty$  is the zero vector (of  $\mathbb{R}^{m+1+p}$ ). Using the calculus rule III.2.2.5, we see in summary that  $M(\bar{x})$  is nonempty and bounded if and only if

$$H_\infty \cap K_\infty = \{0\}. \quad (2.4.5)$$

We have  $H_\infty = \mathbb{R}^m \times \{0\} \times \mathbb{R}^p$ , while  $K_\infty = K$  (we have seen that  $K$  is a nonempty closed convex cone). In other words, (2.4.5) means

$$[\mathbb{R}^m \times \{0\} \times \mathbb{R}^p] \cap [\{0\} \cup M_0(\bar{x})] = \{0\},$$

and this is exactly (2.4.4). □

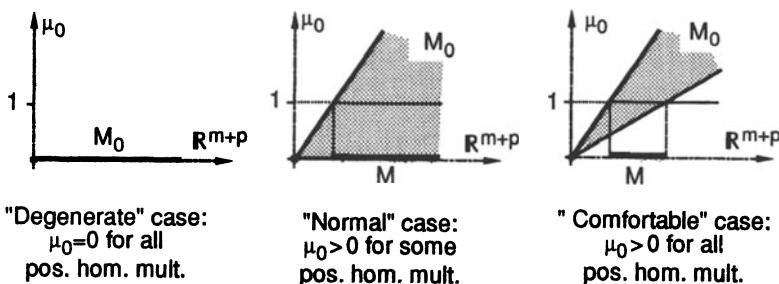


Fig. 2.4.1. Different possibilities for the multipliers

Note that (2.4.4) can be expressed in the following equivalent form: the only  $(\lambda, \mu) \in \mathbb{R}^m \times (\mathbb{R}^+)^p$  satisfying

$$0 \in \sum_{i=1}^m \lambda_i a_i + \sum_{j \in J(\bar{x})} \mu_j \partial c_j(\bar{x})$$

is the zero vector. Naturally, this is only another form of strong Slater assumption: take  $\mu = 0$  to obtain the linear independence of the  $a_i$ 's; and realize that  $\bar{x}$  cannot minimize the function  $\max_j c_j$  under the constraint  $Ax = b$ . Thus, if the property holds at some  $\bar{x} \in C$ , it holds throughout  $C$ .

Figure 2.4.1 displays the various possibilities concerning the sets of multipliers; it also illustrates the proof of Theorem 2.4.5. Finally, Fig. 2.4.2 summarizes the main results of §2.

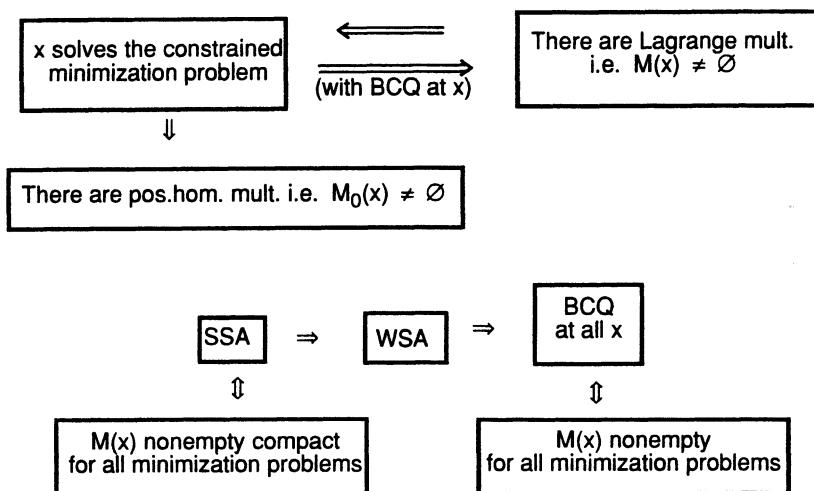


Fig. 2.4.2. Connection between minimality conditions and Qualification Conditions

### 3 Properties and Interpretations of the Multipliers

#### 3.1 Multipliers as a Means to Eliminate Constraints: the Lagrange Function

From their very definition in §2, the multipliers seem to depend on the data  $(f, a, b, c)$  of the constrained minimization problem (2.0.3), and also on the particular solution  $\bar{x}$  considered. Actually, they do not depend on the latter.

**Proposition 3.1.1** *Let  $\bar{x}$  and  $\bar{x}'$  be two solutions of the constrained minimization problem (2.0.3),  $M(\bar{x})$  and  $M(\bar{x}')$  being their associated sets of Lagrange multipliers. Then  $M(\bar{x}) = M(\bar{x}')$ .*

PROOF. By definition,  $(\lambda, \mu) \in M(\bar{x})$  means

$$0 \in \partial f(\bar{x}) + A^* \lambda + \sum_{j=1}^p \mu_j \partial c_j(\bar{x}), \quad (*)$$

$$\mu_j \geq 0 \text{ and } \mu_j c_j(\bar{x}) = 0 \quad \text{for } i = 1, \dots, p. \quad (**)$$

Consider the convex function

$$\mathbb{R}^n \ni x \mapsto \ell_{\lambda, \mu}(x) := f(x) + \lambda^\top(Ax - b) + \sum_{j=1}^p \mu_j c_j(x)$$

and observe that (\*) just expresses the fact that  $\bar{x}$  is an unconstrained minimum of  $\ell_{\lambda, \mu}$ . Thus, we can write  $\ell_{\lambda, \mu}(\bar{x}') \geq \ell_{\lambda, \mu}(\bar{x})$ . Straightforward calculations using (\*\*) yield

$$f(\bar{x}') + \sum_{j=1}^p \mu_j c_j(\bar{x}') \geq f(\bar{x}) = \bar{f} = f(\bar{x}').$$

Because of the signs of  $\mu_j$  and  $c_j(\bar{x}')$ , this implies

$$\mu_j c_j(\bar{x}') = 0 \quad \text{for all } j = 1, \dots, p. \quad (3.1.1)$$

It follows

$$\ell_{\lambda, \mu}(\bar{x}') = f(\bar{x}') = f(\bar{x}) = \ell_{\lambda, \mu}(\bar{x}),$$

i.e.  $\bar{x}'$  is an unconstrained minimum of  $\ell_{\lambda, \mu}$ : (\*) holds with  $\bar{x}$  replaced by  $\bar{x}'$ . Together with (3.1.1), we finally have  $(\lambda, \mu) \in M(\bar{x}')$ .

Thus, we have proved  $M(\bar{x}) \subset M(\bar{x}')$ ; the converse inclusion follows by symmetry.  $\square$

The same proof could have been used to establish  $M_0(\bar{x}) = M_0(\bar{x}')$ . However, from now on, we will pay attention to the Lagrange multipliers exclusively.

**Remark 3.1.2** As a result of Proposition 3.1.1, we are entitled to use the notation  $M$  for the set  $M(\bar{x})$  of Lagrange multiplier. This notation is symmetric to the notation  $S$  for the solution-set of (2.0.3).

In particular, the  $\mu$ -part of the Lagrange multipliers does not depend on the solution  $\bar{x}$ ; and because of the complementarity slackness (see Remark 2.1.5),  $\mu_j$  has to be zero for  $j = 1, \dots, p$ , as soon as there is a solution  $\bar{x}$  with  $c_j(\bar{x}) < 0$ . Expressed otherwise,

$$\{j = 1, \dots, p : \mu_j > 0\} \subset \cap\{J(\bar{x}) : \bar{x} \in S\}.$$

Because the set  $J(\bar{x})$  of active constraints does depend on  $\bar{x}$ , we say: when  $S$  increases, the chances of having the strict complementarity slackness decrease.  $\square$

A by-product of the proof of Proposition 3.1.1 is a function  $\ell_{\lambda, \mu}$ , which is minimal at  $\bar{x}$ . This function is in fact fundamental in optimization and deserves a formal definition:

**Definition 3.1.3** The *Lagrange function*, or *Lagrangian*, associated with the constrained minimization problem (2.0.3) is the function  $L : \mathbb{R}^n \times \mathbb{R}^{m+p} \rightarrow \mathbb{R}$  defined by

$$(x, \lambda, \mu) \mapsto L(x, \lambda, \mu) := f(x) + \sum_{i=1}^m \lambda_i ((a_i, x) - b_i) + \sum_{j=1}^p \mu_j c_j(x).$$

Using the notation from (2.0.2), we will also write more simply

$$L(x, \lambda, \mu) = f(x) + \lambda^\top(Ax - b) + \mu^\top c(x). \quad \square$$

It is important to understand that  $L$  is a “bivariate” function: it depends on the two groups of variables  $x \in \mathbb{R}^n$  and  $(\lambda, \mu) \in \mathbb{R}^{m+p}$ , which play quite distinct roles. For fixed  $(\lambda, \mu)$ ,  $L(\cdot, \lambda, \mu)$  is a *convex* function from  $\mathbb{R}^n$  to  $\mathbb{R}$ ; and the “interesting” values of the variable  $x$  form the set  $S$ . Alternatively, for fixed  $x$ ,  $L(x, \cdot, \cdot)$  is *affine* on  $\mathbb{R}^{m+p}$ ; and the “interesting” values of the variable  $(\lambda, \mu)$  form the set  $M \subset \mathbb{R}^m \times (\mathbb{R}^+)^p$ .

From Proposition 3.1.1,  $M$  is an intrinsic object, attached to the data  $(f, a, b, c)$  of our constrained minimization problem (2.0.3). The next result goes along the same lines: by contrast to the definition (2.1.8), (2.1.9) of  $M$  (which implies to have  $\bar{x}$  first), it establishes the ability of  $M$  to produce optimal solutions via an unconstrained minimization problem.

**Proposition 3.1.4** *For  $(\lambda, \mu) \in M$ , the two statements below are equivalent:*

- (i)  $\bar{x} \in C$  minimizes  $L(\cdot, \lambda, \mu)$  over  $\mathbb{R}^n$  and  $\mu_j c_j(\bar{x}) = 0$  for  $j = 1, \dots, p$ ;
- (ii)  $\bar{x}$  solves the original problem (2.0.3).

PROOF. It suffices to observe that (i) is just the KKT conditions (2.1.8), (2.1.9), with an equivalent formulation of (2.1.8).  $\square$

As a first illustration, let us return to the steepest-descent problem of Example 2.3.5. To obtain a solution, we could minimize on  $\mathbb{R}^n$  the convex function

$$d \mapsto L(d, \mu) := \langle s, d \rangle + \frac{1}{2}\mu(Qd, d) - \frac{1}{2}\mu\kappa$$

for some multiplier  $\mu$ , which turned out to be positive. The constant term  $\frac{1}{2}\mu\kappa$  plays no role and can be dropped; if, furthermore,  $\mu Q = \nabla^2 f(x)$  (assuming existence of the Hessian), we recognize in the minimand the second-order approximation of  $f(x + d) - f(x)$ . This confirms that Newtonian methods can be considered as steepest-descent methods with a suitable norming of  $\mathbb{R}^n$ .

The end of Remark II.2.1.4 can also be illustrated: as an alternative to (2.3.3), consider

$$\min \left\{ \frac{1}{2}\langle Qd, d \rangle : \langle s, d \rangle = \delta \right\}.$$

It has just one affine constraint, so it is equivalent to minimizing for some  $\lambda$  the function

$$d \mapsto \frac{1}{2}\langle Qd, d \rangle + \lambda\langle s, d \rangle,$$

which gives  $d = -\lambda Q^{-1}s$ . The multiplier is then obtained from the feasibility condition:  $\lambda = -\delta/\langle s, Q^{-1}s \rangle$ . Once again,  $d$  depends multiplicatively on  $\lambda$ , i.e. on  $\delta$ , and equivalence with (2.3.3) follows if  $\lambda < 0$ , i.e. if  $\delta < 0$  (a sensible requirement, since a positive value of  $\delta$  would result in an uphill direction).

Let us sum up the information furnished by a multiplier  $(\lambda, \mu) \in M$ :

- The values in two minimization problems are equal, namely:

$$\inf \{f(x) : x \in C\} = \inf \{L(x, \lambda, \mu) : x \in \mathbb{R}^n\}$$

(as can be seen from the proof of Proposition 3.1.1).

- The solutions of the first problem are those  $\bar{x}$  solving the second problem which are in  $C$  and satisfy the complementarity slackness:  $c_j(\bar{x}) = 0$  if  $\mu_j > 0$ .

This aspect will be further developed in §4 below.

### 3.2 Multipliers and Exact Penalty

In §1.2, we introduced another way of eliminating constraints, through penalty. Our study there was based on the distance-function  $d_C$ ; but this function can usually not be computed with the help of the problem-data  $(a, b, c)$  of (2.0.1). On the other hand, we saw that  $d_C$  satisfied the nice *exact penalty* property of Definition 1.2.2: it reproduced a solution of the original problem, provided that it was amplified by a large enough penalty coefficient  $\pi$ .

Here let us introduce more formally a penalty function, depending explicitly on the data  $(a, b, c)$ . For  $i = 1, \dots, m$  and  $j = 1, \dots, p$ , we choose individual penalty functions  $p_i, q_j$ , all convex from  $\mathbb{R}$  to  $\mathbb{R}^+$ , satisfying the following properties:

$$\begin{aligned} p_i(0) &= 0 \quad \text{and} \quad p_i(t) > 0 \text{ for } t \neq 0; \\ q_j(t) &= 0 \text{ for } t \leq 0 \quad \text{and} \quad q_j(t) > 0 \text{ for } t > 0. \end{aligned} \quad (3.2.1)$$

Then we construct

$$\mathbb{R}^n \ni x \mapsto P(x) := \sum_{i=1}^m p_i(\langle a_i, x \rangle - b_i) + \sum_{j=1}^p q_j(c_j(x)); \quad (3.2.2)$$

this  $P$  is a penalty function: it satisfies (1.2.1).

Recall our substitute unconstrained problem:

$$\inf \{f(x) + P(x) : x \in \mathbb{R}^n\}. \quad (3.2.3)$$

We know from Lemma 1.2.1 that a solution of (3.2.3) solves the original constrained problem as soon as it is feasible; and to force this feasibility, each  $p_i(t)$  and  $q_j(t)$  should increase fast enough when  $t$  leaves 0: remember Fig. 1.2.1. The wording “fast enough” is precisely made clear by the corresponding Lagrange multiplier:

**Lemma 3.2.1** *Let  $(\lambda, \mu) \in M$ . If the penalty functions (3.2.1) are chosen so that*

$$\lambda_i \in \partial p_i(0) \quad \text{for } i = 1, \dots, m, \quad (3.2.4)$$

$$\mu_j \in \partial q_j(0) \quad \text{for } j = 1, \dots, p, \quad (3.2.5)$$

*then any solution of the original problem (2.0.3) solves the penalized problem (3.2.3).*

PROOF. For each  $i = 1, \dots, m$ , the subgradient inequality (3.2.4) gives:

$$[p_i(0) + \lambda_i t] = \lambda_i t \leq p_i(t) \quad \text{for all } t \in \mathbb{R}.$$

Taking successively  $t = \langle a_i, x \rangle - b_i$  for  $i = 1, \dots, m$  and summing up, we obtain

$$\lambda^\top (Ax - b) \leq \sum_{i=1}^m p_i(\langle a_i, x \rangle - b_i) \quad \text{for all } x \in \mathbb{R}^n.$$

Starting from (3.2.5), we similarly arrive at

$$\mu^\top c(x) \leq \sum_{j=1}^p q_j(c_j(x)) \quad \text{for all } x \in \mathbb{R}^n$$

and we deduce by summation:

$$\lambda^\top(Ax - b) + \mu^\top c(x) \leq P(x) \quad \text{for all } x \in \mathbb{R}^n. \quad (3.2.6)$$

Now let  $\bar{x}$  solve (2.0.3) and write

$$\begin{aligned} f(\bar{x}) + P(\bar{x}) &= f(\bar{x}) \\ &= f(\bar{x}) + \lambda^\top(A\bar{x} - b) + \mu^\top c(\bar{x}) && [\text{feasibility of } \bar{x}] \\ &\leq f(x) + \lambda^\top(Ax - b) + \mu^\top c(x) && [\text{with transversality}] \\ &\leq f(x) + P(x) && [\text{Proposition 3.1.4}] \\ &\leq f(x) + P(x) && [\text{using (3.2.6)}] \end{aligned}$$

In a word,  $\bar{x}$  solves (3.2.3).  $\square$

**Remark 3.2.2** It is instructive to note that the above proof is almost identical to that of Proposition 3.1.4. Indeed, with  $(\lambda, \mu) \in M$ , the Lagrange function  $L(\cdot, \lambda, \mu)$  resembles, *but is not*, a penalized function, in which each of the individual  $p_i$  and  $q_j$  would be linear, with slopes  $\lambda_i$  and  $\mu_j$  respectively. The geometrical meaning of  $\lambda, \mu$  appears in Fig. 3.2.1.  $\square$

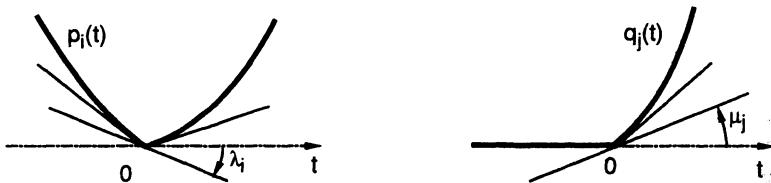


Fig. 3.2.1. Behaviour of a penalty term near zero

Now, remembering Lemma 1.2.1, the above result makes it easy to find an implementable exact penalty function:

**Corollary 3.2.3** *Let the constrained minimization problem (2.0.3) have a nonempty (solution-set and) set of Lagrange multipliers. Then the function  $\gamma_1$  of (2.0.8) satisfies the exact penalty property of Definition 1.2.2. More precisely, if*

$$\pi > \pi^* := \max \{|\lambda_1|, \dots, |\lambda_m|; \mu_1, \dots, \mu_p\},$$

*then the solutions of (2.0.3) are the unconstrained minima of  $f + \pi \gamma_1$ .*

**PROOF.** The function  $q := \pi \gamma_1$  has the form described via (3.2.1), (3.2.2), with  $p_i(t) = \pi |t|$  and  $q_j(t) = \pi t^+$ . If  $\pi$  is as stated, we see from Lemma 3.2.1 that  $q$  satisfies the conditions of Lemma 1.2.1(iii).  $\square$

Having thus singled out an exact penalty function, we can obtain many others; a simple variant has one penalty coefficient for each constraint:

$$p_i(t) = \pi_i |t| \text{ for } i = 1, \dots, m \quad \text{and} \quad q_j(t) = \pi_{m+j} t^+ \text{ for } j = 1, \dots, p,$$

where  $\pi_i > |\lambda_i|$ ,  $\pi_{m+j} > \mu_j$ . Actually, consider the vector-function  $\Gamma$  of (2.0.5): because all norms are equivalent, any function of the type  $\|\Gamma\|$  is an exact penalty when  $M \neq \emptyset$ . Another useful example is (with  $\pi > 0$ )

$$p_i(t) = \pi |\exp t - 1|, \quad q_j(t) = \pi \max\{\exp t - 1, 0\}.$$

These functions tend to  $+\infty$  rapidly when  $t \rightarrow +\infty$ ; therefore, they increase the chances of having compact sublevel-sets in (3.2.3).

Thus, the property  $M \neq \emptyset$  suffices to provide a convenient exact penalty function. This condition turns out to be necessary, which is not surprising in view of Remark 3.2.2.

**Theorem 3.2.4** *For the constrained minimization problem (2.0.3) (assumed to have a solution), the three statements below are equivalent:*

- (i)  $M \neq \emptyset$ ;
- (ii) there is a penalty function  $P$  of the form (3.2.2) such that any solution of the original problem (2.0.3) solves the penalized problem (3.2.3);
- (iii) there is a penalty function  $P$  of the form (3.2.2) such that the original problem (2.0.3) and the penalized problem (3.2.3) have the same solution-set.

PROOF. (i)  $\Rightarrow$  (ii) is Lemma 3.2.1; (ii)  $\Leftrightarrow$  (iii) is Corollary 3.2.3, the only thing to prove is (ii)  $\Rightarrow$  (i).

Let  $\bar{x}$  solve (2.0.3) – hence (3.2.3) – and write the minimality condition  $0 \in \partial(f + P)(\bar{x})$ : from the appropriate calculus rules of §VI.4,

$$0 \in \partial f(\bar{x}) + \sum_{i=1}^m \partial p_i(0) a_i + \sum_{j=1}^p \partial q_j(c_j(\bar{x})) \partial c_j(\bar{x}).$$

This displays  $\lambda_i \in \partial p_i(0)$ ,  $i = 1, \dots, m$  and  $\mu_j \in \partial q_j(c_j(\bar{x}))$ ,  $j = 1, \dots, p$  such that (2.1.8) holds. It remains only to check (2.1.9), which follows easily from the properties (3.2.1) of  $q_j$ :  $[\mu_j \in] \partial q_j(t) \subset \mathbb{R}^+$  for all  $t$  and  $[\{\mu_j\} =] \partial q_j(t) = \{0\}$  if  $t < 0$ .  $\square$

So far, we have seen three possibilities to remove the constraints from the original problem (2.0.3):

- (i) A penalty function can be added to  $f$ ; in practice, it depends on a penalty coefficient  $\pi$ , which is increased until the penalized problem (hopefully) produces a feasible solution.
- (ii) Linear functions of the constraints can be added to  $f$ , thus forming the Lagrange function; the coefficients  $\lambda, \mu$  of these linear functions must be adjusted so that a minimum is (hopefully) obtained, which is feasible and satisfies the complementarity slackness.

- (iii) A shift  $\bar{f}$  can be subtracted from  $f$ , to form the max-function  $F$  of (2.4.1), which must be minimized on the affine manifold of equation  $Ax = b$ . The value  $\bar{f}$  (here viewed as an unknown parameter) must be adjusted so that a minimizer  $x^*$  of  $F$  is obtained, (hopefully) satisfying  $F(x^*) = f(x^*) - \bar{f} = 0$ .

None of these approaches is straightforward, in the sense that they all contain some unknown parameter:  $\pi$ ,  $(\lambda, \mu)$ , or  $\bar{f}$ . The techniques (i) and (ii) require the existence of a multiplier, i.e. in practice the weak Slater assumption; by contrast, (iii) is satisfied merely with the existence of a solution. On the other hand, (i) appears as the most tractable:  $\pi$  just has to be large enough; while (ii) and (iii) require an accurate value of their respective parameters.

At any rate, most methods for constrained minimization possess an interpretation in terms of at least one of the techniques (i), (ii), (iii).

### 3.3 Multipliers as Sensitivity Parameters with Respect to Perturbations

For many applications, it is important to study the behaviour of the optimal value of a minimization problem such as (2.0.3), when the data  $(f, a, b, c)$  vary. We consider here perturbations in the right-hand sides of the constraints only: other perturbations result in much more involved studies, which would go beyond the scope of this book; incidentally, the behaviour of the optimal solutions is likewise a delicate subject.

Thus, for  $(u, v) = (u_1, \dots, u_m; v_1, \dots, v_p) \in \mathbb{R}^m \times \mathbb{R}^p$ , we consider

$$\left| \begin{array}{l} \inf f(x) \\ \langle a_i, x \rangle - b_i = u_i \text{ for } i = 1, \dots, m \quad [\text{or } Ax = b + u \in \mathbb{R}^m] \\ c_j(x) \leq v_j \text{ for } j = 1, \dots, p. \quad [\text{or } v - c(x) \in (\mathbb{R}^+)^p] \end{array} \right. \quad (3.3.1)_{u,v}$$

Of course,  $(3.3.1)_{0,0}$  is simply our original problem (2.0.3). We call  $C(u, v)$  the feasible set in  $(3.3.1)_{u,v}$  and  $P(u, v)$  the optimal value:

$$P(u, v) := \inf \{f(x) : x \in C(u, v)\}.$$

We still assume that the original problem does have a solution  $\bar{x}$ , i.e.

$$P(0, 0) = \inf \{f(x) : x \in C(0, 0)\} = f(\bar{x}), \quad \text{with } \bar{x} \in C(0, 0);$$

but anything can happen for  $(u, v) \neq (0, 0)$ , even close to  $(0, 0)$ :  $C(u, v)$  may be empty (then  $P(u, v) = +\infty$  by convention), or  $f$  may not be bounded from below on  $C(u, v)$ , i.e.  $P(u, v) = -\infty$ . Thus  $P$  assumes its values a priori in  $\mathbb{R} \cup \{\pm\infty\}$ . Note, however, the following property of  $P$ :

$$v \leq v' \text{ componentwise in } \mathbb{R}^p \implies P(u, v) \geq P(u, v'). \quad (3.3.2)$$

In this section, we consider the following questions:

- When is the value  $-\infty$  excluded for  $P(u, v)$ ? (from §IV.2.4,  $P$  will then be in  $\text{Conv } \mathbb{R}^{m+p}$ ).

- When is  $P$  finite in a neighborhood of  $(0,0)$ ? (from Theorem IV.3.1.2,  $P$  will then be Lipschitzian near the origin).
- At what speed does  $P(u, v)$  tend to  $P(0, 0)$ ? (this depends on the subdifferential of  $P$  at the origin).
- What are the constraints provoking the largest perturbation on  $P$ ?

The answers to all these questions lie in the set of multipliers at  $(3.3.1)_{0,0}$ . To get an idea of what can be expected and what is hopeless, we start with an example.

**Example 3.3.1** Consider again Example 2.4.4:

$$C(v) = \{x = (\xi, \eta) \in \mathbb{R}^2 : \xi + \|\xi, \eta\| \leq v\}.$$

If  $v < 0$ ,  $C(v) = \emptyset$ . If  $v > 0$ , direct calculations give the parabolic set of Fig. 3.3.1:

$$C(v) = \{(\xi, \eta) : 2\xi \leq v - \eta^2/v\}.$$

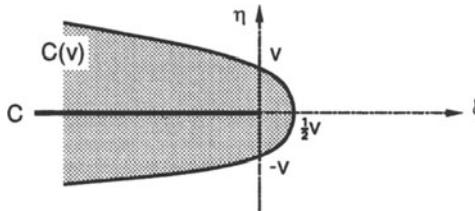


Fig. 3.3.1. A parabolic constraint-set

With the objective function  $f_\alpha(\xi, \eta) = \eta - \alpha\xi$  ( $\alpha \geq 0$ ), we have  $P(0) = 0$ ,  $P(v) = +\infty$  if  $v < 0$ ; as for  $v > 0$ , there are two cases:

- if  $\alpha = 0$ , then  $P(v) = -\infty$ .
- if  $\alpha > 0$ , solving the perturbed problem is a good exercise to apply §2; we find

$$\bar{x}(v) = \left(\frac{1}{2}v - \frac{1}{2\alpha^2}v, -\frac{1}{\alpha}v\right), \quad P(v) = -\frac{1}{2}v\left(\alpha + \frac{1}{\alpha}\right).$$

In this example, we recall that there is no Lagrange multiplier if  $\alpha = 0$ ; and for  $\alpha > 0$ , the set of multipliers is nonempty but unbounded.  $\square$

Clearly enough, if  $P$  is finite in a neighborhood of  $(0, 0)$ , then the strong Slater assumption has to hold: indeed,

- if  $A$  is not surjective, then a perturbation  $(u, 0)$  with  $u$  close to 0 but out of  $\text{Im } A$  will make  $C(u, 0)$  empty;
- if  $\max_j c_j(x) \geq 0$  for all  $x$  satisfying  $Ax = b$ , a perturbation  $(0, v)$  with all components of  $v$  negative will again make  $C(0, v)$  empty.

**Theorem 3.3.2** *The set  $M$  (possibly empty) of multipliers associated with the original minimization problem  $(3.3.1)_{0,0}$  is the set of  $(\lambda, \mu)$  for which it holds that*

$$P(u, v) \geq P(0, 0) - \lambda^\top u - \mu^\top v \quad \text{for all } (u, v) \in \mathbb{R}^m \times \mathbb{R}^p. \quad (3.3.3)$$

*It follows that, if  $M \neq \emptyset$ , then  $P$  assumes nowhere the value  $-\infty$  and is therefore in  $\text{Conv } \mathbb{R}^{m+p}$ .*

PROOF. Let  $(\lambda, \mu) \in M$ . We know from Proposition 3.1.4 that

$$P(0, 0) \leq L(x, \lambda, \mu) \quad \text{for all } x \in \mathbb{R}^n.$$

In particular, we deduce for all  $x \in C(u, v)$ :

$$\begin{aligned} P(0, 0) - \lambda^\top u - \mu^\top v &\leq L(x, \lambda, \mu) - \lambda^\top u - \mu^\top v \\ &= f(x) + \lambda^\top(Ax - b - u) + \mu^\top[c(x) - v] \\ &= f(x) + \mu^\top[c(x) - v] \leq f(x), \end{aligned}$$

and (3.3.3) follows since  $x$  was arbitrary in  $C(u, v)$ . Existence of such a  $(\lambda, \mu)$  therefore implies  $P(u, v) > -\infty$  for all  $(u, v)$ : by virtue of Corollary IV.2.4.3,  $P \in \text{Conv } \mathbb{R}^{m+p}$  (remember that we have assumed  $P(0, 0) = f(\bar{x}) < +\infty$  from the very beginning).

Conversely, suppose (3.3.3) holds and let  $x$  be arbitrary in  $\mathbb{R}^n$ ; taking  $u = Ax - b$  and  $v = c(x)$ , we have

$$f(x) \geq P(Ax - b, c(x)) \geq P(0, 0) - \lambda^\top(Ax - b) - \mu^\top c(x), \quad (3.3.4)$$

where the first inequality holds because  $x$  is certainly in  $C(Ax - b, c(x))$ . This can be written

$$P(0, 0) \leq f(x) + \lambda^\top(Ax - b) + \mu^\top c(x) = L(x, \lambda, \mu),$$

hence  $P(0, 0) = f(\bar{x}) = \min L(\cdot, \lambda, \mu)$ .

Now set  $x = \bar{x}$  in (3.3.4) to get

$$f(\bar{x}) = P(0, 0) \geq P(0, 0) - \mu^\top c(\bar{x}).$$

It remains only to establish nonnegativity of  $\mu$ , but this results from the monotonicity property (3.3.2): take in (3.3.3)  $u = 0$  and  $v = e_j$ , the  $j^{\text{th}}$  vector of the canonical basis in  $\mathbb{R}^p$ .  $\square$

Note the illustration of (3.3.3) given by Example 3.3.1 (the set of multipliers was computed in Example 2.4.4). In the same example, take the objective function  $f(\xi, \eta) = e^\eta - 1$ . When  $v \leq 0$ , nothing is changed; but now,  $P(v) = -1 > -\infty$  for  $v > 0$ : existence of multipliers is not necessary to exclude the value  $-\infty$  for  $P$ . Needless to say,  $M$  is still empty and the new  $P$  has no slope at 0.

We now turn to the property  $P(u, v) < +\infty$  near  $(0, 0)$ . Once again, remember that we postulate the existence of an optimal solution to the unperturbed problem.

**Theorem 3.3.3** *The strong Slater assumption is necessary and sufficient for  $P$  to be finite in a neighborhood of  $(0, 0)$ .*

PROOF. [Sufficiency] Because SSA implies  $M \neq \emptyset$ , we already know that the value  $-\infty$  is excluded. Let  $x_0$  satisfy

$$Ax_0 = b \quad \text{and} \quad c_j(x_0) \leq -\varepsilon \quad \text{for } j = 1, \dots, p$$

for some  $\varepsilon > 0$ . Take a perturbation  $v$  such that

$$|v|_\infty := \max \{|v_j| : j = 1, \dots, p\} \leq \varepsilon/2.$$

- Because each  $c_j$  is continuous, there is a ball  $B(x_0, r)$  around  $x_0$  such that, for all  $x \in B(x_0, r)$  and  $j = 1, \dots, p$

$$c_j(x) \leq c_j(x_0) + \varepsilon/2 \leq -\varepsilon/2 \leq -|v_j| \leq v_j.$$

- Because  $A$  is surjective,  $AA^*$  is invertible (see Example 1.1.5). For a perturbation  $u \in \mathbb{R}^m$ , the vector  $x_u := A^*(AA^*)^{-1}u + x_0$  satisfies  $Ax_u = b + u$  and furthermore,  $x_u \in B(x_0, r)$  if  $u$  is close enough to 0.

Thus, for  $(u, v)$  close enough to 0,  $C(u, v)$  is nonempty, hence  $P(u, v) < +\infty$ .

[Necessity] This property was already alluded to after Example 3.3.1, we give one more proof. When  $P$  is finite in a neighborhood of  $(0, 0)$ , i.e. when  $(0, 0) \in \text{int dom } P$ , we know from §VI.1 that the subdifferential  $\partial P(0, 0)$  is a nonempty compact convex set of  $\mathbb{R}^{m+p}$ . On the other hand, Theorem 3.3.2 tells us that this set is exactly  $-M$ ; the rest follows from Theorem 2.3.2.  $\square$

Thus, under SSA, the equality  $-M = \partial P(0, 0)$  allows the following refinement of (3.3.3):

$$P(u, v) = P(0, 0) + \sigma_M(-u, -v) + o(\|(u, v)\|). \quad (3.3.5)$$

It follows that  $P$  is differentiable at  $(0, 0)$  if and only if there is exactly one multiplier.

In this case, when some  $u_i$  becomes nonzero, the optimal cost is perturbed to  $P(0, 0) - \lambda_i u_i$ , up to first order. Thus,  $\lambda_i$  represents a “value” of the  $i^{\text{th}}$  constraint: to keep satisfying it, one is ready to pay a price  $\lambda_i u_i$  balancing the above perturbation. This interpretation also explains that, to find an optimal solution of the original problem, this  $i^{\text{th}}$  constraint must be assigned a price  $\lambda_i$ ; we recover Proposition 3.1.4.

**Example 3.3.4** Take again the example of minimizing a quadratic function over an affine manifold:

$$P(b + u) = \min \left\{ \frac{1}{2} \langle Qx, x \rangle + \langle c, x \rangle : Ax = b + u \right\}, \quad (3.3.6)_u$$

where  $Q$  is symmetric positive definite and  $A$  is surjective. Using from Example 1.1.5 the expression of  $\bar{x}$  solving  $(3.3.6)_0$ , straightforward computations give

$$P(b + u) = P(b) + \langle BAQ^{-1}c + Bb, u \rangle + \frac{1}{2} \langle Bu, u \rangle.$$

The optimal value is thus a quadratic function of the right-hand side; its gradient at  $u = 0$  is  $B(AQ^{-1}c + b)$ , in which we recognize the expression of  $-\lambda$  in Example 1.1.5.  $\square$

**Example 3.3.5** Consider an infimal convolution: using the notation of §VI.4.5, we solve the constrained minimization problem in  $\mathbb{R}^{2n}$

$$(f_1 \downarrow f_2)(x) := \inf \{f_1(y_1) + f_2(y_2) : y_1 + y_2 = x\}$$

for given  $x \in \mathbb{R}^n$ . Recalling that we write the constraint as  $A(y_1, y_2) = x$ ,  $A$  is surjective from  $\mathbb{R}^{2n}$  to  $\mathbb{R}^n$ : SSA is satisfied. If the infimal convolution is exact at  $x = y_1 + y_2$ , i.e. if the above problem has a solution  $(y_1, y_2)$ , we know that there is

a nonempty bounded set of multipliers, whose opposite is exactly the subdifferential of  $f_1 \downarrow f_2$  at  $x$ . Indeed the minimality conditions give the multipliers, namely those  $\lambda$  satisfying

$$0 \in \partial f_1(y_1) \times \partial f_2(y_2) + (\lambda, \lambda);$$

this is Corollary VI.4.5.5.  $\square$

## 4 Minimality Conditions and Saddle-Points

The study of our basic convex minimization problem (2.0.3) has revealed two sets which play different roles: one is the set  $S \subset \mathbb{R}^n$  of *solutions* to (2.0.3) (assumed nonempty in our development); and the second is the set  $M \subset \mathbb{R}^m \times (\mathbb{R}^+)^p$  of *multipliers* (whose nonemptiness is guaranteed under appropriate assumptions). All our work of §2 has consisted in associating to any  $\bar{x} \in S$  the elements  $(\lambda, \mu)$  of  $M$ . Conversely, for given  $(\lambda, \mu) \in M$ , we have seen in §3.1 how to obtain the solution-set  $S$ . The two groups of variables,  $x \in \mathbb{R}^n$  and  $(\lambda, \mu) \in \mathbb{R}^m \times (\mathbb{R}^+)^p$ , are gathered in a “bivariate” function, the Lagrangian.

In the present section, the multipliers  $(\lambda, \mu)$  will be shown to solve a concave maximization problem associated with (2.0.3): the *dual problem*. The product  $S \times M$  will appear as something new: the set of *saddle-points*, forming the solutions of a “mixed” problem of *mini-maximization*.

We make an elementary study, directly inspired from the previous section: here are our first steps in duality theory. Deeper and more detailed analysis will come in Chap. XII, which is entirely devoted to the subject; in particular, the role of convexity will be demonstrated there, as well as the consequences of this theory on decomposition aspects.

We start with general considerations on the extremization of bivariate functions.

### 4.1 Saddle-Points: Definitions and First Properties

Given a function  $f : X \subset \mathbb{R}^n \rightarrow \mathbb{R}$ , we know the meaning of expressions like “minimize  $f$  on  $X$ ”, “minimal value of  $f$  on  $X$ ”, “minimum-set of  $f$  on  $X$ ”; and symmetric concepts are obtained if “min” is replaced by “max”. We now consider optimization problems of another type, concerning a “bivariate” function  $\ell$ , which depends on two distinct groups of variables  $x$  and  $y$ .

Let  $X$  and  $Y$  be two nonempty sets and consider a given function  $\ell$

$$X \times Y \ni (x, y) \mapsto \ell(x, y) \in \mathbb{R}.$$

Suppose that we want to minimize  $\ell$  with respect to  $x$ , and maximize it with respect to  $y$ . For each value of one variable, the set (possibly empty) of extremizers in the other variable is then relevant, say:

$$\begin{aligned} T(x) &:= \left\{ \tilde{y} \in Y : \ell(x, \tilde{y}) = \sup_{y \in Y} \ell(x, y) \right\}, \\ S(y) &:= \left\{ \tilde{x} \in X : \ell(\tilde{x}, y) = \inf_{x \in X} \ell(x, y) \right\}. \end{aligned}$$

This defines two multifunctions,  $T : X \rightarrow Y$  and  $S : Y \rightarrow X$ , whose graphs are subsets of  $X \times Y$  and  $Y \times X$  respectively.

**Definition 4.1.1** A couple  $(\bar{x}, \bar{y}) \in X \times Y$  is said to be a *saddle-point* of  $\ell$  on  $X \times Y$  when

$$\bar{y} \in T(\bar{x}) \quad \text{and} \quad \bar{x} \in S(\bar{y}). \quad (4.1.1)$$

□

Just from the definition of upper and lower bounds, a saddle-point is a couple  $(\bar{x}, \bar{y})$  such that

$$\sup_{y \in Y} \ell(\bar{x}, y) = \ell(\bar{x}, \bar{y}) = \inf_{x \in X} \ell(x, \bar{y}), \quad (4.1.2)$$

where the sup and inf must actually be a min and a max; this in turn can be written

$$\ell(\bar{x}, y) \leq \ell(\bar{x}, \bar{y}) \leq \ell(x, \bar{y}) \quad \text{for all } (x, y) \in X \times Y. \quad (4.1.3)$$

A further definition is

$$\ell(\bar{x}, y) \leq \ell(x, \bar{y}) \quad \text{for all } (x, y) \in X \times Y. \quad (4.1.4)$$

Indeed, (4.1.4) is clearly implied by (4.1.3); conversely, if (4.1.4) holds, take successively  $y = \bar{y}$  and  $x = \bar{x}$  to obtain (4.1.3). As we continue, we will use indifferently one of the possible definitions (4.1.1) – (4.1.4) – and even several others, to be seen below.

For a saddle-point  $(\bar{x}, \bar{y})$ , we have in particular  $\bar{x} \in \text{dom } T$  and  $\bar{y} \in \text{dom } S$ ; indeed, (4.1.1) can also be written  $(\bar{x}, \bar{y}) \in \text{gr } T$  and  $(\bar{y}, \bar{x}) \in \text{gr } S$ . In a set-theoretic language, the (possibly empty) set of saddle-points is  $\text{gr } T \cap \text{gr}^T S$ , where

$$\text{gr}^T S := \{(x, y) \in X \times Y : (y, x) \in \text{gr } S\}$$

is the “symmetrized” version of  $\text{gr } S$ .

Let us start with simple examples.

**Example 4.1.2** With  $X = Y = \mathbb{R}$ , the graph of the quadratic form defined by

$$\ell(x, y) = x^2 - y^2$$

resembles a saddle, depicted on Fig. 4.1.1. Another comparison is topographical: we have two mountains on the  $x$ -axis, separated by a pass at the point  $(0, 0)$ . When moving from  $(0, 0)$  in the  $x$ -direction [ $y$ -direction],  $\ell$  increases [decreases];  $(0, 0)$  appears as the only saddle-point of  $\ell$ .

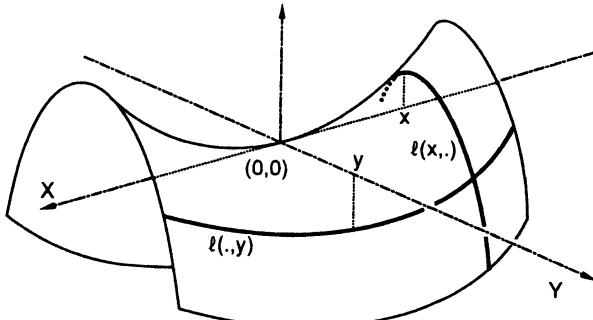
A more subtle example is the quadratic form

$$\mathbb{R} \times \mathbb{R}^+ \ni (x, y) \mapsto \ell(x, y) = xy - x. \quad (4.1.5)$$

Here, the multifunctions  $T$  and  $S$  are given by

$$T(x) = \begin{cases} \emptyset & \text{if } x > 0, \\ \mathbb{R}^+ & \text{if } x = 0, \\ \{0\} & \text{if } x < 0, \end{cases} \quad S(y) = \begin{cases} \emptyset & \text{if } y \neq 1, \\ \mathbb{R} & \text{if } y = 1, \end{cases}$$

and  $(0, 1)$  is the saddle-point of  $\ell$  on  $\mathbb{R} \times \mathbb{R}^+$ . We will see in §4.4 that this example is typical in the framework of the present chapter. □



**Fig. 4.1.1.** A saddle-shaped function

Saddle-points form a set in  $X \times Y$  which is not quite arbitrary. It has the structure of a Cartesian product, say  $\bar{S} \times \bar{T}$  with  $\bar{S} \subset X$  and  $\bar{T} \subset Y$ :

**Proposition 4.1.3** *The value  $\ell(\bar{x}, \bar{y})$  is constant over all saddle-points  $(\bar{x}, \bar{y})$ . If  $(\bar{x}_1, \bar{y}_1)$  and  $(\bar{x}_2, \bar{y}_2)$  are two saddle-points of  $\ell$  on  $X \times Y$ , then so are  $(\bar{x}_1, \bar{y}_2)$  and  $(\bar{x}_2, \bar{y}_1)$ .*

PROOF. We have by the definition (4.1.3) of saddle-points:

$$\text{for all } (x, y) \in X \times Y, \begin{cases} \ell(\bar{x}_1, y) \leq \ell(\bar{x}_1, \bar{y}_1) \leq \ell(x, \bar{y}_1), \\ \ell(\bar{x}_2, y) \leq \ell(\bar{x}_2, \bar{y}_2) \leq \ell(x, \bar{y}_2). \end{cases} \quad (*)$$

Set  $x = \bar{x}_2$  and  $y = \bar{y}_2$  in  $(*)$ ,  $x = \bar{x}_1$  and  $y = \bar{y}_1$  in  $(**)$  to see that the four values  $\ell(\bar{x}_i, \bar{y}_j)$  are equal for  $i$  and  $j$  in  $\{1, 2\}$ .

Then use successively  $(*)$  and  $(**)$ : for all  $(x, y) \in X \times Y$ ,

$$\ell(\bar{x}_1, y) \leq \ell(\bar{x}_1, \bar{y}_1) = \ell(\bar{x}_2, \bar{y}_2) \leq \ell(x, \bar{y}_2),$$

hence  $(\bar{x}_1, \bar{y}_2)$  is a saddle-point; play the same trick for  $(\bar{x}_2, \bar{y}_1)$ .  $\square$

The real number  $\bar{\ell} := \ell(\bar{x}, \bar{y})$  singled out by this result is called the *saddle-value*. In our topographical interpretation of Fig. 4.1.1,  $\bar{\ell}$  is the altitude of the pass. The following is then one more possible definition:  $(\bar{x}, \bar{y})$  is a saddle-point when there is a number  $\bar{\ell}$  such that

$$\ell(\bar{x}, y) \leq \bar{\ell} \leq \ell(x, \bar{y}) \quad \text{for all } (x, y) \in X \times Y. \quad (4.1.6)$$

**Remark 4.1.4** If  $(\bar{x}, \bar{y})$  is a saddle-point, we have by definition (4.1.1)  $\bar{x} \in S(T(\bar{x}))$ , i.e.  $\bar{x} \in (S \circ T)(\bar{x})$ ; and likewise  $\bar{y} \in (T \circ S)(\bar{y})$ . We can say that  $\bar{x}$  and  $\bar{y}$  are *fixed points* of the multifunctions  $S \circ T: X \longrightarrow X$  and  $T \circ S: Y \longrightarrow Y$  respectively.

Conversely suppose  $\bar{x}$  is a fixed point of  $S \circ T$ : there is  $\tilde{y} \in T(\bar{x})$  such that  $\bar{x} \in S(\tilde{y})$ . By definition,  $(\bar{x}, \bar{y})$  is a saddle-point:  $\bar{x}$  is in the  $\bar{S}$ -part of the set of saddle-points; symmetrically, a fixed point  $\bar{y}$  of  $T \circ S$  is in the  $\bar{T}$ -part of this same set. In view of the Cartesian structure assessed in Proposition 4.1.3, we therefore obtain a further characterization of saddle-points: they are those pairs  $(\bar{x}, \bar{y})$  whose components are fixed points of  $S \circ T$  and  $T \circ S$  respectively.  $\square$

## 4.2 Mini-Maximization Problems

The definition (4.1.1) of a saddle-point involved the pair of (set-valued) mappings  $T$ =“Argmax” and  $S$ =“Argmin”. To reach the equivalent definition (4.1.2), we could directly use function-values. Following this idea, we define two functions:

$$\begin{aligned} X \ni x &\mapsto \varphi(x) := \sup_{y \in Y} \ell(x, y), \\ Y \ni y &\mapsto \psi(y) := \inf_{x \in X} \ell(x, y); \end{aligned} \quad (4.2.1)$$

note that  $\varphi$  can take on the value  $+\infty$ , and  $\psi$  the value  $-\infty$ , even in very simple situations: see (4.1.5) for example.

**Lemma 4.2.1** *For  $\varphi$  and  $\psi$  defined by (4.2.1), there holds*

$$\psi(y) \leq \ell(x, y) \leq \varphi(x) \quad \text{for all } (x, y) \in X \times Y. \quad (4.2.2)$$

PROOF. Take  $x$  and  $y$  in  $X$  and  $Y$  respectively. By definition of an inf:

$$\psi(y) \leq \ell(x, y),$$

and of a sup:

$$\ell(x, y) \leq \varphi(x).$$

□

This result reveals an important general property: all  $\psi$ -values *minorize* all  $\varphi$ -values; to memorize it, think that when one minimizes, one obtains something smaller than when one maximizes. The whole business in finding a saddle-point is then to invert the inequalities in (4.2.2):

**Proposition 4.2.2** *With  $\varphi$  and  $\psi$  defined by (4.2.1),  $(\bar{x}, \bar{y})$  is a saddle-point of  $\ell$  on  $X \times Y$  if and only if*

$$\psi(\bar{y}) \geq \varphi(\bar{x}). \quad (4.2.3)$$

*Then we actually have  $\psi(\bar{y}) = \varphi(\bar{x}) = \ell(\bar{x}, \bar{y}) =: \bar{\ell}$ .*

PROOF. Because of Lemma 4.2.1, (4.2.3) means exactly

$$\psi(\bar{y}) = \ell(\bar{x}, \bar{y}) = \varphi(\bar{x}),$$

i.e., by definition of  $\varphi$  and  $\psi$ :

$$\inf_{x \in X} \ell(x, \bar{y}) = \ell(\bar{x}, \bar{y}) = \sup_{y \in Y} \ell(\bar{x}, y),$$

which is just (4.1.2). □

See Fig. 4.2.1, where both  $T(\cdot)$  and  $S(\cdot)$  are assumed single-valued throughout; a saddle-point is obtained when the points  $A$ ,  $A_-$  and  $A^+$  coincide (we are then on the intersection of the curves  $\text{gr } S$  and  $\text{gr } T$ ). In Fig. 4.1.1,  $\text{gr } S$  and  $\text{gr } T$  are the two coordinate-axes.

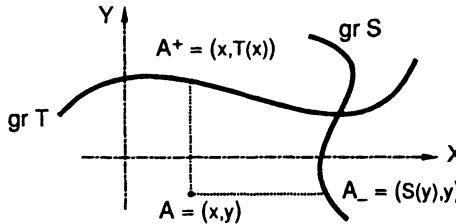


Fig. 4.2.1. Mini-maximization

Compare (4.2.2) with (4.2.3) to realize that finding a saddle-point implies maximizing the inf-function  $\psi$ , and minimizing the sup-function  $\varphi$ . Geometrically, embed Fig. 4.2.1 in  $\mathbb{R}^3$ , with the extra component  $z = \ell(x, y)$ : we must find a point on the curve  $\text{gr } T[\text{gr } S]$  with highest [lowest] altitude  $z$ . Naturally, (4.2.2) extends to the extremal values:

$$\sup_{y \in Y} \inf_{x \in X} \ell(x, y) \leq \inf_{x \in X} \sup_{y \in Y} \ell(x, y), \quad (4.2.4)$$

a relation which holds for all  $(X, Y, \ell)$ . Proposition 4.2.2 implies that, when there is a saddle-point, *equality holds* in (4.2.4): both sides are then equal to the saddle-value  $\bar{\ell}$ . Such an equality need not hold in general, however:

**Example 4.2.3** For  $x = (\xi, \eta) \in X = \mathbb{R}^2$  and  $y \in Y = \mathbb{R}^+$ , take

$$\ell(x, y) = y + y \left[ \xi + \sqrt{\xi^2 + \eta^2} \right].$$

This function is convex in  $x$ . Observe that  $\partial_x \ell$  never contains 0: no matter how  $y$  is chosen,  $\ell(\cdot, y)$  has no minimum; as a result, there cannot be any saddle-point:  $S(y) = \emptyset$  for all  $y$ .

Indeed, we have  $\psi(y) \equiv -\infty$  (take for example  $\xi = y\eta$  and  $\eta \rightarrow -\infty$ ); the left-hand side in (4.2.4) is therefore  $-\infty$ . On the other hand,

$$\varphi(\xi, \eta) = \begin{cases} 0 & \text{if } \xi + \sqrt{\xi^2 + \eta^2} = 0, \\ +\infty & \text{otherwise.} \end{cases}$$

In summary, (4.2.4) gives  $-\infty \leq 0$ . □

The trouble illustrated by this example is that the operators “min” and “max” do not commute in general. This means in Fig. 4.2.1 that the two curves  $\text{gr } S$  and  $\text{gr } T$  have no reason a priori to meet each other. Existence of a saddle-point implies above all that the order of extremizations of  $\ell$  does not matter: one can start equivalently with  $x$ , or with  $y$ .

**Remark 4.2.4** When (4.2.4) holds as an equality, we can say that  $\ell$  has a saddle-value. Beware that this property does not imply the existence of a saddle-point: each of the outer extremizations must still have a solution. A simple counter-example is

$$[1, +\infty[ \times [1, +\infty[ \ni (x, y) \mapsto \ell(x, y) = \frac{1}{x} - \frac{1}{y}.$$

This example illustrates another interesting point: suppose that, instead of  $X = Y = [1, +\infty[$ , we take

$$X = [1, L] \quad \text{and} \quad Y = [1, L']$$

for some real numbers  $L$  and  $L'$  (larger than 1). On these new sets,  $\ell$  has the saddle-point  $(L, L')$ , with its saddle-value  $1/L - 1/L'$ . When  $L$  and  $L'$  tend to infinity, this saddle-point has no limit, but the saddle-value tends to 0. □

Knowing that the function  $\varphi$  [resp.  $\psi$ ] of (4.2.1) *must* be minimized [maximized], how about the converse? To obtain a saddle-point, does it *suffice* to extremize these functions? Consider their respective sets (possibly empty) of extremizers:

$$\Phi := \{x^* \in X : \varphi(x^*) = \inf \varphi\}, \quad \Psi := \{y^* \in Y : \psi(y^*) = \sup \psi\}.$$

Computing  $\Phi$  is a mini-maximization problem:  $\ell$  is first maximized (with respect to  $y$ ), and then minimized (with respect to  $x$ ) – and the other way round for  $\Psi$ . The precise connection between saddle-points and this type of hierarchical problems lies in the following result, which also specifies the sets  $\bar{S}$  and  $\bar{T}$  appearing in Proposition 4.1.3.

**Theorem 4.2.5** *With the above notation, a necessary and sufficient condition for  $\ell$  to have a saddle-point on  $X \times Y$  is*

$$\min_{x \in X} \varphi(x) = \max_{y \in Y} \psi(y); \quad (4.2.5)$$

in this case, the set of saddle-points is  $\Phi \times \Psi$  ( $\neq \emptyset$ ).

PROOF. If  $(\bar{x}, \bar{y})$  is a saddle-point, it is obvious from (4.1.2) that  $(\bar{x}, \bar{y}) \in \Phi \times \Psi$ , and that  $\varphi(\bar{x}) = \ell(\bar{x}, \bar{y}) = \psi(\bar{y})$ : (4.2.5) holds.

Conversely, assume (4.2.5) and take  $(\bar{x}, \bar{y}) \in \Phi \times \Psi$ : we have  $\psi(\bar{y}) = \varphi(\bar{x})$ . Then, by definition of  $\varphi$  and  $\psi$ ,

$$\ell(\bar{x}, y) \leq \varphi(\bar{x}) = \psi(\bar{y}) \leq \ell(x, \bar{y}) \quad \text{for all } (x, y) \in X \times Y;$$

by virtue of (4.1.4),  $(\bar{x}, \bar{y})$  is a saddle-point. □

Note the following implication of this (fundamental) result: when (4.2.5) holds, not only do  $\varphi$  and  $\psi$  attain their extrema; but also, at any of these extrema, say  $\bar{x}$  and  $\bar{y}$ , the functions  $\ell(\bar{x}, \cdot)$  and  $\ell(\cdot, \bar{y})$  also attain their extrema. Because nothing else is presupposed from  $(X, Y, \ell)$ , we conclude that (4.2.5) must be a very strong property, indeed.

**Remark 4.2.6** Let us sum up this subsection: at least conceptually, the computation of a saddle-point can be done as follows:

- (i) For each  $x \in X$ , the function  $\varphi(x)$  of (4.2.1) must be computed;
- (ii) minimize  $\varphi$  to obtain a minimum point  $\bar{x}$  (if none exists, there is no saddle-point);
- (iii) having  $\bar{x}$ , maximize  $\ell(\bar{x}, \cdot)$  to obtain a maximum  $y^*$  (if none exists, there is no saddle-point).
- (i') Likewise, for each  $y \in Y$ , the function  $\psi(y)$  of (4.2.1) must be computed;
- (ii') maximize  $\psi$  to obtain a maximum point  $\bar{y}$  (if none exists, there is no saddle-point);
- (iii') having  $\bar{y}$ , minimize  $\ell(\cdot, \bar{y})$  to obtain a minimum  $x^*$  (if none exists, there is no saddle-point).

When these computations are completed, compare  $\ell(\bar{x}, y^*)$  with  $\ell(x^*, \bar{y})$  [i.e.  $\varphi(\bar{x})$  with  $\psi(\bar{y})$ ]. If they are equal,  $(\bar{x}, \bar{y})$  is a saddle-point; otherwise, there is no saddle-point.

Beware that the points  $x^* \in S(\bar{y})$  and  $y^* \in T(\bar{x})$  obtained in (iii') and (iii) may have nothing to do with what we are looking for. It is only when  $x^*$  and  $y^*$  are uniquely determined that they form a saddle-point; this will be shown in the existence Theorem 4.3.1 below.  $\square$

Thus, three problems can be considered, associated with  $(X, Y, \ell)$ : the saddle-point problem of Definition 4.1.1, the *minimax* problem

$$\min_{x \in X} \varphi(x) \quad \text{i.e.} \quad \min \left\{ \max_{y \in Y} \ell(x, y) : x \in X \right\}, \quad (4.2.6)$$

and the *maximin* problem

$$\max_{y \in Y} \psi(y) \quad \text{i.e.} \quad \max \left\{ \min_{x \in X} \ell(x, y) : y \in Y \right\}.$$

They are related, but they can coincide only if  $(X, Y, \ell)$  enjoys rather strong properties. Finally, we introduce the abbreviated notation for (4.2.6):

$$\min_{x \in X} \max_{y \in Y} \ell(x, y). \quad (4.2.7)$$

### 4.3 An Existence Result

The previous sections have given some general properties of saddle-points and minimaximization problems in an abstract setting; but we still know nothing about existence of a solution. Theorem 4.2.5, however, gives an idea of the type of assumptions needed:

(i) First, equality must hold in (4.2.4), and this property is far from automatic. Take for example the case illustrated by Fig. 4.3.1:  $X = Y = \{0, 1\}$ , and  $\ell(0, 0) = 1$ ,  $\ell(1, 1) = 2$ ,  $\ell(1, 0) = 3$ ,  $\ell(0, 1) = 4$ . We have  $\varphi(0) = 4$ ,  $\varphi(1) = 3$ , whose minimal value is 3; and  $\psi(0) = 1$ ,  $\psi(1) = 2$ ; thus (4.2.4) gives here  $2 \leq 3$ .

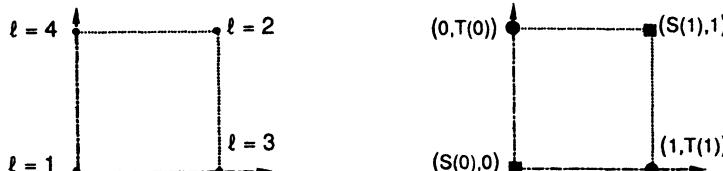


Fig. 4.3.1. No saddle-value

Not surprisingly, since we are dealing with extremization problems, convexity must come into play. Convenient assumptions are:

$$X \subset \mathbb{R}^n \text{ and } Y \subset \mathbb{R}^m \text{ are nonempty closed convex sets} \quad (H1)$$

$$\left. \begin{array}{l} \ell \text{ is convex-concave on } X \times Y \text{ in the following sense:} \\ \text{for each } y \in Y, \text{ the function } \ell(\cdot, y) : X \rightarrow \mathbb{R} \text{ is convex,} \\ \text{for each } x \in X, \text{ the function } \ell(x, \cdot) : Y \rightarrow \mathbb{R} \text{ is concave.} \end{array} \right\} \quad (H2)$$

- (ii) Second, when minimizing  $\ell$  with respect to  $x$  (for some specific  $y$ ), we must obtain a solution; remembering Remark IV.3.2.6, what is needed here is 0-coercivity of the function  $\ell(\cdot, y) + I_X$ ; and likewise for  $y$ . Thus, we also assume

$$\left. \begin{array}{l} X \text{ is bounded, or there exists } y_0 \in Y \text{ such that} \\ \ell(x, y_0) \rightarrow +\infty \text{ when } \|x\| \rightarrow +\infty, x \in X; \end{array} \right\} \quad (\text{H3})$$

$$\left. \begin{array}{l} Y \text{ is bounded, or there exists } x_0 \in X \text{ such that} \\ \ell(x_0, y) \rightarrow -\infty \text{ when } \|y\| \rightarrow +\infty, y \in Y. \end{array} \right\} \quad (\text{H4})$$

All these assumptions are symmetric and natural; they imply in particular the continuity of  $\ell$  with respect to each of its two arguments. We mention here that the convexity Assumptions (H1), (H2) do not suffice to ensure equality in (4.2.4): see Example 4.2.3. Some compactness is still missing, and (H3), (H4) do the job.

**Theorem 4.3.1** *Under Assumptions (H1) – (H4),  $\ell$  has a nonempty convex compact set of saddle-points on  $X \times Y$ .*

PROOF. First of all, we know from Proposition 4.1.3 that the set of saddle-points, when nonempty, has the form  $\bar{S} \times \bar{T}$ ; from (4.1.6),  $\bar{S}$  is an intersection of sublevel-sets of a convex function; it is closed and convex. From (H3), either all these sublevel-sets are bounded, or at least one of them is bounded (the one corresponding to  $y_0$ ). In both cases, their intersection  $\bar{S}$  is bounded. The same argument establishes convexity and compactness of  $\bar{T}$ .

For non-emptiness, we proceed in three steps: first, we prove existence under additional assumptions; then we remove these assumptions one after the other.

[Step 1] In addition to (H1) – (H4), assume that

$$X \text{ and } Y \text{ are bounded, } \ell(x, \cdot) \text{ is strictly concave for each } x \in X. \quad (4.3.1)$$

Consider the family of functions indexed by  $y \in Y$ :

$$\mathbb{R}^n \ni x \mapsto \ell(x, y) + I_X(x);$$

They are in  $\text{Conv} \mathbb{R}^n$ , and our assumptions imply the following properties:

- they attain their maximum for each  $x \in X$ , at a unique  $y = T(x)$ ;
- by Theorem IV.2.1.2, the resulting max-function  $\varphi$  of (4.2.1) is in  $\text{Conv} \mathbb{R}^n$ ;
- the domain of  $\varphi$  is the compact set  $X$ , and  $\varphi$  attains its minimum at some  $\bar{x} \in X$ .

Now let  $x$  be arbitrary in  $X$  and, for  $k = 1, 2, \dots$ , define

$$x_k := \frac{1}{k}x + (1 - \frac{1}{k})\bar{x} \quad \text{and} \quad y_k := T(x_k).$$

Applying successively the definition of  $\bar{x}$  and of  $y_k$ , and the convexity of  $\ell(\cdot, y_k)$ , we have

$$\varphi(\bar{x}) \leq \varphi(x_k) = \ell(x_k, y_k) \leq \frac{1}{k}\ell(x, y_k) + (1 - \frac{1}{k})\ell(\bar{x}, y_k). \quad (4.3.2)$$

Let  $k \rightarrow +\infty$ ; because of (4.3.1),  $\{y_k\}$  has a cluster point, say  $\bar{y}$ . Passing to the limit in (4.3.2),

$$\varphi(\bar{x}) \leqslant 0 + \ell(\bar{x}, \bar{y}),$$

so  $\bar{y} = T(\bar{x})$  is *independent* of  $x$ .

We claim that  $(\bar{x}, \bar{y})$  is a saddle-point of  $\ell$ . We already have

$$\ell(\bar{x}, \bar{y}) = \varphi(\bar{x}) \geqslant \ell(\bar{x}, y) \quad \text{for all } y \in Y,$$

which is the first half of (4.1.3). On the other hand, use the definition of  $\varphi$  in (4.3.2) to obtain

$$\varphi(\bar{x}) \leqslant \frac{1}{k} \ell(x, y_k) + (1 - \frac{1}{k}) \varphi(\bar{x});$$

multiply by  $k$ :

$$\varphi(\bar{x}) \leqslant \ell(x, y_k)$$

and let  $y_k \rightarrow \bar{y}$  to obtain the second half of (4.1.3): we do have a saddle-point of  $\ell$ .

[Step 2] Now, in addition to (H1) – (H4), assume only that  $X$  and  $Y$  are bounded. For  $k = 1, 2, \dots$  consider the function

$$X \times Y \ni (x, y) \mapsto \ell_k(x, y) := \ell(x, y) - \frac{1}{k} \|y\|^2$$

which is strictly concave in  $y$ . From the first step above,  $\ell_k$  has a saddle-point, say  $(\bar{x}_k, \bar{y}_k)$ : for all  $(x, y) \in X \times Y$ ,

$$\ell(\bar{x}_k, y) - \frac{1}{k} \|y\|^2 \leqslant \ell(x, \bar{y}_k) - \frac{1}{k} \|\bar{y}_k\|^2.$$

Let  $k \rightarrow +\infty$ , extract a subsequence if necessary so that  $(\bar{x}_k, \bar{y}_k) \rightarrow (\bar{x}, \bar{y}) \in X \times Y$  and pass to the limit:

$$\ell(\bar{x}, y) \leqslant \ell(x, \bar{y}) \quad \text{for all } (x, y) \in X \times Y,$$

which is (4.1.4):  $(\bar{x}, \bar{y})$  is a saddle-point of  $\ell$ .

[Step 3] Finally, just assume (H1) – (H4) and, for  $k = 1, 2, \dots$ , define the two convex compact sets  $X_k := X \cap B(0, k)$ ,  $Y_k := Y \cap B(0, k)$ . Then, from Step 2,  $\ell$  has a saddle-point  $(\bar{x}_k, \bar{y}_k)$  on  $X_k \times Y_k$ :

$$\ell(\bar{x}_k, y) \leqslant \ell(x, \bar{y}_k) \quad \text{for all } (x, y) \in X_k \times Y_k. \quad (4.3.3)$$

Let  $k \rightarrow +\infty$  and suppose that  $\{\bar{y}_k\}$  is unbounded; then  $Y$  is unbounded as well so, for  $k$  large enough,  $X_k$  contains the point  $x_0$  of (H4); using it in (4.3.3),

$$\ell(\bar{x}_k, y) \leqslant \ell(x_0, \bar{y}_k) \rightarrow -\infty.$$

Thus,  $\ell(\bar{x}_k, y) \rightarrow -\infty$ , which can happen only for an unbounded  $\{\bar{x}_k\}$ ; this, however, is impossible: reasoning as above, we exhibit  $y_0 \in Y_k$  from (H3) and use it in (4.3.3) to obtain the contradiction

$$+\infty \leftarrow \ell(\bar{x}_k, y_0) \leqslant \ell(x_0, \bar{y}_k) \rightarrow -\infty.$$

The same proof establishes boundedness of  $\{\bar{x}_k\}$ . Then, we take a cluster-point of  $\{(\bar{x}_k, \bar{y}_k)\}$ ; to show that it is a saddle-point of  $\ell$  on  $X \times Y$ , we proceed as in Step 2, passing to the limit in (4.3.3).  $\square$

It goes without saying that, if  $\ell(\cdot, y)$  is strictly convex for all  $y \in Y$  [resp.  $\ell(x, \cdot)$  is strictly concave for all  $x \in X$ ], then the set of saddle-points has the form  $\{\bar{x}\} \times \bar{T}$  [resp.  $\bar{S} \times \{\bar{y}\}$ ].

**Remark 4.3.2** Use the results of §1.1: under the convexity assumptions (H1), (H2), a saddle-point  $(\bar{x}, \bar{y})$  is characterized by the existence of  $\bar{s} \in \partial_x \ell(\bar{x}, \bar{y})$  and  $\bar{p} \in \partial_y(-\ell)(\bar{x}, \bar{y})$  such that

$$\begin{aligned}\langle \bar{s}, x - \bar{x} \rangle &\geq 0 \quad \text{for all } x \in X, \\ \langle \bar{p}, y - \bar{y} \rangle &\geq 0 \quad \text{for all } y \in Y.\end{aligned}$$

A consequence of Theorem 4.3.1 is that this system has a solution when (H3), (H4) hold.  $\square$

#### 4.4 Saddle-Points of Lagrange Functions

After the general study of Sections 4.1 to 4.3, let us consider the particular case where  $\ell$  is the Lagrange function of Definition 3.1.3, associated with the basic minimization problem (2.0.3). Thus, the context is now the following:

$$\begin{aligned}X &= \mathbb{R}^n \quad \text{with variable } x \text{ as before,} \\ Y &= \mathbb{R}^m \times (\mathbb{R}^+)^p \quad \text{with variable } y = (\lambda, \mu), \\ \ell &= L : \mathbb{R}^n \times \mathbb{R}^m \times (\mathbb{R}^+)^p \rightarrow \mathbb{R} \quad \text{is defined by} \\ L(x, \lambda, \mu) &:= f(x) + \lambda^\top(Ax - b) + \mu^\top c(x),\end{aligned}$$

and we are interested in saddle-points of  $L$ . With respect to the previous subsections, the essential simplification concerns the  $(\lambda, \mu)$ -variable:  $L$  is affine, and its maximization is made over a closed convex cone. We know (Example 1.1.6) the characterization of such maxima; the definition (4.1.1) of a saddle-point can be reformulated accordingly:

**Proposition 4.4.1** *With the system of notation as above, the saddle-points of  $L$  on  $\mathbb{R}^n \times [\mathbb{R}^m \times (\mathbb{R}^+)^p]$  are those  $(\bar{x}, (\bar{\lambda}, \bar{\mu}))$  such that*

- (i)  $\bar{x}$  minimizes  $L(\cdot, \bar{\lambda}, \bar{\mu})$  on  $\mathbb{R}^n$ ;
- (ii)  $A\bar{x} = b$  and  $c_j(\bar{x}) \leq 0$  for  $j = 1, \dots, p$  [ $\bar{x}$  is feasible in (2.0.3)];
- (iii)  $\bar{\mu}_j c_j(\bar{x}) = 0$  for  $j = 1, \dots, p$  [transversality condition].

PROOF. (i) is nothing but the second half of (4.1.2).

As for the first half of (4.1.2), it expresses that  $(\bar{\lambda}, \bar{\mu})$  solves the optimization problem

$$\max \{L(\bar{x}, \lambda, \mu) : (\lambda, \mu) \in \mathbb{R}^m \times (\mathbb{R}^+)^p\}.$$

The optimality condition of Theorem 1.1.1 can be worked out; but the situation is actually much simpler, since  $L(\bar{x}, \cdot, \cdot)$  can be maximized separately with respect to each  $\lambda_i$  and each  $\mu_j$ . A solution  $(\bar{\lambda}, \bar{\mu})$  is then clearly characterized as follows:

- for  $i = 1, \dots, m$ , the slope  $\langle a_i, \bar{x} \rangle - b_i$  is zero;
- for  $j = 1, \dots, p$ , the slope  $c_j(\bar{x})$  is
 
$$\begin{cases} \text{zero if } \bar{\mu}_j > 0, \\ \text{nonpositive if } \bar{\mu}_j = 0. \end{cases}$$

Altogether, we recover (ii), (iii).  $\square$

The above proof illustrates the difficulty mentioned after Remark 4.2.6. If  $\bar{x}$  answers the question,  $L(\bar{x}, \cdot, \cdot)$  is maximized on a huge set: for example, any  $\lambda \in \mathbb{R}^m$  is optimal. Only a small part of this set, however, is likely to answer the question.

**Corollary 4.4.2** *If  $(\bar{x}, (\bar{\lambda}, \bar{\mu}))$  is a saddle-point of  $L$  over  $\mathbb{R}^n \times [\mathbb{R}^m \times (\mathbb{R}^+)^p]$ , then  $\bar{x}$  solves the basic minimization problem (2.0.3).*

PROOF. If  $(\bar{x}, (\bar{\lambda}, \bar{\mu}))$  is a saddle-point, (i), (ii), (iii) hold in Proposition 4.4.1;  $\bar{x}$  is feasible. Furthermore, (iii), (i) give

$$f(\bar{x}) = L(\bar{x}, \bar{\lambda}, \bar{\mu}) \leq L(x, \bar{\lambda}, \bar{\mu}) \quad \text{for all } x \in \mathbb{R}^n;$$

then, it suffices to observe that  $\bar{\mu} \in (\mathbb{R}^+)^p$  implies

$$L(x, \bar{\lambda}, \bar{\mu}) \leq f(x) \quad \text{for all feasible } x. \quad \square$$

The sufficiency condition thus obtained makes the link between the present Section 4 and the general scope of this chapter. It is worth noting that the above two results are very general: no assumption whatsoever is needed on the data  $(f, a, b, c)$ ; the equality constraints could even be non-affine. However, as mentioned already, existence of a saddle-point of  $L$  is a very restrictive property; as indicated by §4.3, convexity plays its role for that:

**Theorem 4.4.3** *In the basic minimization problem (2.0.3), assume that the data  $f$  and  $c_j$ ,  $j = 1, \dots, p$  are convex functions from  $\mathbb{R}^n$  to  $\mathbb{R}$ . Then the two statements below are equivalent:*

- (i)  $(\bar{x}, (\bar{\lambda}, \bar{\mu}))$  is a saddle-point of  $L$  over  $\mathbb{R}^n \times [\mathbb{R}^m \times (\mathbb{R}^+)^p]$ ;
- (ii)  $\bar{x}$  solves (2.0.3) and  $(\bar{\lambda}, \bar{\mu})$  is a Lagrange multiplier.

PROOF. Use Proposition 4.4.1 and the minimality condition  $0 \in \partial_x L(\bar{x}, \bar{\lambda}, \bar{\mu})$ . Then recognize in  $(\bar{\lambda}, \bar{\mu})$  the definition of Lagrange multipliers, as given by Theorem 2.1.4(iv).  $\square$

Remember Proposition 3.1.1, and also Proposition 4.1.3 establishing that the saddle-points form a Cartesian product: the saddle-points of the Lagrange function form the set  $S \times M$ , where  $S$  is the solution-set and  $M$  the set of multipliers. The example of (4.1.5), precisely, is the Lagrange function associated with the following naive constrained minimization problem:

$$\min -x \quad \text{subject to} \quad c(x) := x \leq 0. \quad (4.4.1)$$

Assuming convexity, existence of saddle-points of  $L$  thus amounts to some qualification condition (remember Proposition 2.2.1). The “abnormal” problem studied in Examples 2.1.7 and 2.4.4 just gives Example 4.2.3: no multiplier, hence no saddle-point.

**Remark 4.4.4** Given that our basic constrained minimization problem (2.0.3) is assumed to have a solution  $\bar{x}$ , the existence of Lagrange multipliers is thus equivalent to the existence of a saddle-point of  $L$ . It is then interesting to interpret in this framework the assumptions used by Theorem 4.3.1 to ensure this existence.

Here, (H1) and (H2) are automatic; when (H3) and (H4) hold, the sets  $S$  and  $M$  of solutions and multipliers are both nonempty and bounded. For (2.0.3) to have a nonempty and bounded set of solutions, a classical assumption is the 0-coercivity of  $f + I_C$ :

$$f(x) \rightarrow +\infty \quad \text{if} \quad \|x\| \rightarrow +\infty \quad \text{with } x \in C. \quad (4.4.2)$$

Because  $L(\cdot, \lambda, \mu) \leq f$  on  $C$  for any  $(\lambda, \mu) \in \mathbb{R}^m \times (\mathbb{R}^+)^p$ , (H3) clearly implies (4.4.2). The converse is not true: with  $x = (\xi, \eta) \in \mathbb{R}^2$ , the problem

$$\min |\eta| \quad \text{subject to} \quad \xi = 0 \quad (4.4.3)$$

does satisfy (4.4.2); but there is no real number  $\lambda$  such that

$$L(x, \lambda) = |\eta| + \lambda \xi \rightarrow +\infty \quad \text{when} \quad \|(\xi, \eta)\| \rightarrow +\infty.$$

Now consider (H4); it is incompatible with any equality constraint: use again the counterexample of (4.4.3), with  $(\xi, \eta)$  and  $\lambda$  exchanged. On the other hand, if the problem contains only inequality constraints, (H4) becomes: there exists  $x_0 \in \mathbb{R}^n$  such that

$$[\mu \in (\mathbb{R}^+)^p, \|\mu\| \rightarrow +\infty] \implies L(x_0, \mu) \rightarrow -\infty. \quad (4.4.4)$$

This  $x_0$  is strictly feasible: otherwise, i.e. if  $c_{j_0}(x_0) \geq 0$  for some  $j_0$ , we have  $L(x_0, \mu) \geq f(x_0)$  for  $\mu = te_{j_0}$ ,  $t \rightarrow +\infty$  ( $e_{j_0}$  being the  $j_0^{\text{th}}$  vector of the canonical basis in  $\mathbb{R}^p$ ); (4.4.4) is contradicted.

Thus, (H4) implies the strong Slater assumption. Conversely, if there is  $x_0$  such that  $c_j(x_0) < 0$  for  $j = 1, \dots, p$ , we certainly have (4.4.4). In summary: when there are no equality constraints, (H4) is equivalent to SSA.  $\square$

## 4.5 A First Step into Duality Theory

The previous subsection has connected the optimality conditions of §2 with the saddle-theory of §4.1; let us now apply the results of §4.2 to our Lagrange function. To make it simple, we assume existence of a saddle-point, i.e. (2.0.3) has a solution and the set  $M$  of multipliers is nonempty.

Because  $L(x, \cdot, \cdot)$  is affine, the function  $\varphi$  of (4.2.1) is easy to compute: we obtain

$$\varphi(x) = \begin{cases} f(x) & \text{if } Ax = 0 \text{ and } c_j(x) \leq 0 \text{ for } j = 1, \dots, p \\ +\infty & \text{otherwise;} \end{cases}$$

in other words,  $\varphi = f + I_C$ : minimizing  $\varphi$  amounts to finding a solution  $\bar{x}$  of (2.0.3). At such a solution, the set  $T(\bar{x})$  of (4.1.1) is

$$\mathbb{R}^m \times \{\mu \in (\mathbb{R}^+)^p : \mu_j = 0 \text{ if } c_j(\bar{x}) > 0\}$$

(which does not help much to compute a multiplier).

On the other hand, the function  $\psi$  is now defined by

$$(\lambda, \mu) \mapsto \psi(\lambda, \mu) := \begin{cases} \inf_{x \in \mathbb{R}^n} L(x, \lambda, \mu) & \text{if } (\lambda, \mu) \in \mathbb{R}^m \times (\mathbb{R}^+)^p \\ -\infty & \text{if not.} \end{cases} \quad (4.5.1)$$

**Theorem 4.5.1** *Let the convex minimization problem (2.0.3) have a solution and a nonempty set of Lagrange multipliers. With  $\psi$  defined by (4.5.1), we have*

$$-\psi \in \overline{\text{Conv}}(\mathbb{R}^m \times \mathbb{R}^p),$$

*and the maxima of  $\psi$  are the Lagrange multipliers.*

PROOF. First of all,  $\psi$  is an infimum of affine functions; Theorem 4.2.5 tells us that, by assumption, its maximal value is  $f(\bar{x}) > -\infty$ . It is therefore closed and convex (Proposition IV.2.1.2). Then combine Theorem 4.2.5 with Theorem 4.4.3.  $\square$

In the present context of a Lagrangian, the maximization of  $\psi$  is called the *dual problem*; to complete the terminology, (2.0.3) is then the *primal problem*; and the variables  $(\lambda, \mu)$  are then called the *dual variables*. Here again,  $\psi$  can take the value  $-\infty$ : in (4.4.1),  $-\psi$  is the indicator of  $\{1\}$ .

Because of Theorem 2.3.2, the strong Slater assumption is thus equivalent to  $\psi$  having a nonempty compact set of maximizers. Remembering Remark IV.3.2.6, this in turn means 0-coercivity of  $-\psi$ :

$$\psi(\lambda, \mu) \rightarrow -\infty \quad \text{for } \|(\lambda, \mu)\| \rightarrow +\infty \text{ with } \mu \in (\mathbb{R}^+)^p.$$

In relation with Remark 4.4.4, assume that (2.0.3) has only inequality constraints: 0-coercivity of  $-\psi$  is equivalent to the existence of  $x_0$  such that  $-L(x_0, \cdot)$  is 0-coercive on  $(\mathbb{R}^+)^p$ .

Another remark concerns §3.3: consider the perturbed minimization problem (3.3.1)<sub>u,v</sub>. We have seen that its optimal value  $P(u, v)$  was convex; that SSA was necessary and sufficient for  $P$  to be finite in a neighborhood of  $(0,0)$ ; and also that  $M$  was then the subdifferential  $\partial P(0, 0)$ . Theorem 4.5.1 can be used as an explanation of this last result: call  $\psi_{u,v}$  the dual function associated with the perturbed problem (3.3.1)<sub>u,v</sub>. In a neighborhood of  $(u, v) = (0, 0)$ , SSA still holds and Theorem 4.5.1 tells us that

$$P(u, v) = \max_{\lambda, \mu} \psi_{u,v}(\lambda, \mu) = \max_{\lambda, \mu} [\psi_{0,0}(\lambda, \mu) - u^\top \lambda - v^\top \mu];$$

then  $\partial P(0, 0)$  is given by the calculus rule of §VI.4.4.

Let us sum up the results of this Section 4: the convex minimization problem (2.0.3) can be viewed as a two-stage minimization problem of the type seen in §4.2. To find a multiplier, we must perform the steps (i'), (ii') of Remark 4.2.6. Then, with  $(\bar{\lambda}, \bar{\mu}) \in M$ , minimize  $L(\cdot, \bar{\lambda}, \bar{\mu})$  (on the whole of  $\mathbb{R}^n$ ); what is needed is a minimum point which is feasible and satisfies the complementarity slackness. This last step is facilitated when the Lagrange function is strictly convex in  $x$ : we obtain a unique minimizer, which *has to* solve the original problem (2.0.3).

We conclude with two examples:

**(a) Linear Programming** Suppose that  $\mathbb{R}^n$  is equipped with the standard dot-product, and that (2.0.3) is

$$\left| \begin{array}{l} \inf q^\top x \\ a_j^\top x + b_j \leq 0 \text{ for } j = 1, \dots, p \quad [Ax + b \leq 0 \text{ for short}] \end{array} \right. \quad (4.5.2)$$

where  $q$  and each  $a_j$  are in  $\mathbb{R}^n$ ,  $b = (b_1, \dots, b_p) \in \mathbb{R}^p$ . The Lagrange function

$$L(x, \mu) := (q^\top + \mu^\top A)x + \mu^\top b$$

is easy to minimize in  $x$ . Of course,  $\psi(\mu) = -\infty$  if  $q + A^\top \mu \neq 0$ ; the dual problem is then

$$\sup \{b^\top \mu : \mu \in (\mathbb{R}^+)^p, A^\top \mu + q = 0\}, \quad (4.5.3)$$

another linear program.

Assuming the feasible domain  $C$  nonempty in (4.5.2), WSA holds. Assuming in addition that the objective function is bounded from below on  $C$ , (4.5.2) has a solution (Example V.3.4.5). Under these conditions, there are multipliers, which form the (therefore nonempty) solution-set of (4.5.3); and the two optimal values are equal.

Now we proceed to show that, conversely, existence of a solution in (4.5.3) implies existence of a solution in (4.5.2). For this, we form the Lagrange function associated with (4.5.3):

$$L(\mu, u, v) = -b^\top \mu + u^\top (A^\top \mu + q) - v^\top \mu = (Au - b - v)^\top \mu + u^\top q,$$

which must be minimized for  $\mu \in \mathbb{R}^p$  and maximized for  $(u, v) \in \mathbb{R}^n \times (\mathbb{R}^+)^p$ . The resulting dual function  $\psi(u, v)$  is  $-\infty$  if  $(Au - b - v) \neq 0$ , and  $u^\top q$  otherwise: the dual problem associated with (4.5.3) is therefore

$$\sup \{q^\top u : Au - b = v, v \in (\mathbb{R}^+)^p\}.$$

Setting  $x = -u$ , we obtain exactly (4.5.2).

In summary, a linear program is its own “bidual”; and it has an optimal solution if and only if its dual has an optimal solution as well; in this case, the optimal objective-values coincide in the primal and dual problems.

**Remark 4.5.2** Non-existence of a solution happens in two cases:

- when the objective function is unbounded; then the general result 4.2.1 implies that the dual feasible set *has to be empty*;
- when the feasible set is empty; then the dual problem cannot have an optimal solution: either its objective function is unbounded, or its feasible set is empty (think of an example with  $A = 0$ ).  $\square$

**(b) Quadratic Programming.** Equip  $\mathbb{R}^n$  again with the dot-product (for simplicity) and modify (4.5.2) to

$$\inf \left\{ \frac{1}{2}x^\top Qx + q^\top x : Ax + b \leq 0 \right\}. \quad (4.5.4)$$

Here the  $n \times n$  matrix  $Q$  is symmetric positive definite – another simplifying assumption. This problem has a unique solution if the feasible domain is nonempty. The Lagrange function is

$$L(x, \mu) = \frac{1}{2}x^\top Qx + (q^\top + \mu^\top A)x + \mu^\top b;$$

by assumption, its minimum is attained at the unique

$$x_\mu = -Q^{-1}(q + A^\top \mu).$$

Plugging this value into  $L$ , we obtain the dual problem:

$$\max_{\mu \in (\mathbb{R}^+)^p} \left[ -\frac{1}{2} (q + A^\top \mu)^\top Q^{-1} (q + A^\top \mu) + b^\top \mu \right]. \quad (4.5.5)$$

This is another concave quadratic maximization problem, with a very simple feasible set, but with a possibly degenerate matrix  $AQ^{-1}A^\top$ .

As in the case of linear programming, (4.5.4) satisfies WSA, and has therefore a nonempty set of multipliers: the dual problem (4.5.5) does have a solution. If, in addition,  $A$  is surjective, this solution is unique:  $AQ^{-1}A^\top$  is positive definite (uniqueness can be also seen from the optimality conditions:  $A^\top \mu = -Q\bar{x} - q$  has a unique solution for given  $\bar{x}$ ). Finally, if the constraints in (4.5.4) were equalities, there would be no dual constraints, the dual problem would just be a system of linear equations, an explicit solution would be available; this solution was actually given in Example 1.1.5.

# VIII. Descent Theory for Convex Minimization: The Case of Complete Information

**Prerequisites.** Chapters II, VI; and to a lesser extent: Chap.V (dual norms), Chap. VII (minimality conditions and saddle points).

**Introduction.** In this chapter we begin to study the problem of *computing* a point  $x$  minimizing a convex function  $f$  – as opposed to *characterizing* it, which was the subject of Chap. VII. This computation will be done by an iterative algorithm, of the type exposed in Chap. II. The new feature, of course, is that  $f$  is not supposed to have a gradient  $\nabla f(x)$  varying continuously with  $x$ .

We assume throughout

$$f : \mathbb{R}^n \rightarrow \mathbb{R} \text{ is convex.}$$

This allows us to use the machinery of Chap. VI, expressing the first-order behaviour of a finite-valued convex function, via subdifferentials and directional derivatives.

## 1 Descent Directions and Steepest-Descent Schemes

### 1.1 Basic Definitions

Just as in Chap. II, the methods to be studied here define the next iterate  $x_{k+1}$  from the present one  $x_k$  in two stages: they compute first a *direction of move*  $d_k \in \mathbb{R}^n$ , then a *stepsize*  $t_k > 0$ , and  $x_{k+1}$  is set to  $x_k + t_k d_k$ . They are *descent methods*, in the sense that

$$f(x_{k+1}) < f(x_k)$$

at each iteration; the following is therefore natural:

**Definition 1.1.1** A *descent direction* of the convex function  $f$  at  $x$  is a  $d \in \mathbb{R}^n$  satisfying:

$$\exists t > 0 \quad \text{such that} \quad f(x + td) < f(x).$$

First, we make sure that this definition is consistent with Definition II.2.1.1 (below,  $\sigma_A$  is the support function of the set  $A$ ).

**Theorem 1.1.2** *A descent direction is equivalently defined by any one of the following properties*

$$\left. \begin{array}{l} f'(x, d) < 0; \\ \sigma_{\partial f(x)}(d) < 0; \\ \langle s, d \rangle < 0 \text{ for all } s \in \partial f(x). \end{array} \right\} \quad (1.1.1)$$

PROOF. Everything is easily seen from the various definitions (VI.1.1.1 and VI.1.1.4) of the directional derivative, and from the compactness of  $\partial f(x)$ .  $\square$

Figure 1.1.1 provides an illustration, showing a sublevel-set of  $f$  and the set of descent directions (the interior of the dashed angle) at a non-optimal  $x$ . This set is convex because  $f'(x, \cdot)$  is convex, open because  $f'(x, \cdot)$  is continuous; it is a cone because  $f'(x, \cdot)$  is positively homogeneous. It is the interior of  $T_{Sf(x)}(x)$ , the tangent cone of the sublevel-set of  $f$  at level  $f(x)$ . Unless  $x$  is optimal, this cone is the polar of the cone generated by the subgradients. All this results from §VI.1.3, see in particular Remark VI.1.3.6; the notation  $Sf(x)$  is introduced in (VI.1.3.1).

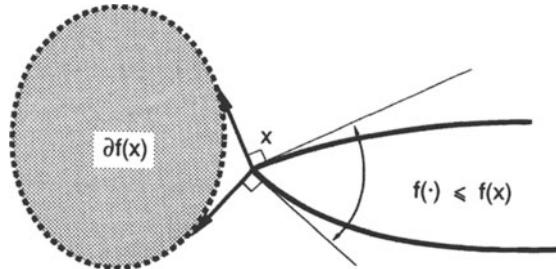


Fig. 1.1.1. The cone of descent directions

When the gradient  $\nabla f(x)$  happens to exist, the subdifferential reduces to the singleton  $\{\nabla f(x)\}$ , which generates a half-line, and the set of descent directions expands to the (open) half-space opposite to  $\nabla f(x)$ . This, in a sense, is the most favourable situation, in which the set of descent directions is as large as possible, just because its polar cone  $\mathbb{R}^+ \nabla f(x)$  is as small as possible.

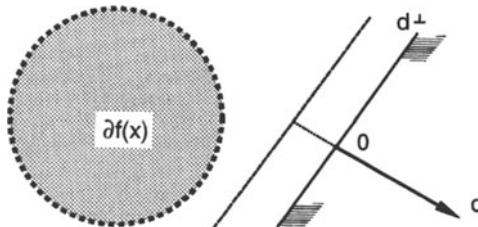


Fig. 1.1.2. Descent directions and separating hyperplanes

Geometrically, a descent direction corresponds to a hyperplane separating the two closed convex sets  $\partial f(x)$  and  $\{0\}$  strictly. Denote by

$$d^\perp := H_{d,0} = \{z \in \mathbb{R}^n : \langle z, d \rangle = 0\}$$

the subspace orthogonal to a given  $0 \neq d \in \mathbb{R}^n$ . Then this  $d$  defines a descent direction when  $\partial f(x)$  lies entirely in the open half-space limited by  $d^\perp$  and opposite to  $d$ . See Fig. 1.1.2 (and compare it with Fig. V.2.1.1); the dashed line, which passes between 0 and  $\partial f(x)$ , is such a separating hyperplane. More precisely:

**Theorem 1.1.3** *A descent direction is a  $d$  such that, if  $\alpha \in [f'(x, d), 0[$  (nonempty by virtue of (1.1.1)), the hyperplane*

$$\{z \in \mathbb{R}^n : \langle z, d \rangle = \alpha\}$$

*separates  $\partial f(x)$  and  $\{0\}$  strictly, i.e.*

$$\langle s, d \rangle \leqslant \alpha < 0 \quad \text{for all } s \in \partial f(x). \quad (1.1.2)$$

PROOF. This is a mere restatement of Theorem 1.1.2. The separation property becomes more obvious if (1.1.2) is rewritten as

$$\langle s, d \rangle \leqslant \alpha < \langle s', d \rangle \quad \text{for all } s \in \partial f(x) \text{ and } s' \in \{0\} \subset \mathbb{R}^n. \quad \square$$

To compute the direction, it is attractive to consider a “best” one, having a directional derivative “as negative as possible”. This concept was already made precise in Definition II.2.1.3, which reads here:

**Definition 1.1.4** *Let  $\|\cdot\|$  be a norm on  $\mathbb{R}^n$ . A normalized steepest-descent direction of  $f$  at  $x$ , associated with  $\|\cdot\|$ , is a solution of the problem*

$$\min \{f'(x, d) : \|d\| = 1\} \quad (1.1.3)$$

or equivalently, using the min-max notation (VII.4.2.7),

$$\min_{\|d\|=1} \max_{s \in \partial f(x)} \langle s, d \rangle. \quad (1.1.4)$$

A non-normalized steepest-descent direction is a  $d \neq 0$  such that  $d/\|d\|$  is a normalized steepest-descent direction.  $\square$

It has already been explained in §II.2.1 that a normalization is necessary, but that the particular value “1” in (1.1.3) is of little importance (see more specifically Fig. II.2.1.1, Remark II.2.1.4). We confirm this fact in the present more general situation where  $f'(x, \cdot)$  is nonlinear: in the following result,  $N$  plays the role of  $\mathbb{R}^n$ ,  $\varphi$  and  $v$  play the role of  $f'(x, \cdot)$  and  $\|\cdot\|$  respectively.

**Proposition 1.1.5** *Let  $N \subset \mathbb{R}^n$  be a cone, let  $\varphi$  and  $v$  be two positively homogeneous functions from  $\mathbb{R}^n$  to  $\mathbb{R}$ . For given  $\kappa > 0$ , call  $D_\kappa$  the set (possibly empty) of solutions of*

$$\min \{\varphi(d) : d \in N, v(d) = \kappa\}. \quad (1.1.5)_\kappa$$

*Then  $D_\kappa = \kappa D_1$  for all  $\kappa > 0$ .*

PROOF. Take  $\lambda > 0$  arbitrary, and suppose that  $d$  solves  $(1.1.5)_\lambda$ . Then  $d \in N$ ,  $v(d) = \lambda$  and

$$\varphi(d') \geq \varphi(d) \quad \text{for all } d' \in N \text{ with } v(d') = \lambda,$$

which can be written (multiply by  $\kappa > 0$  and use positive homogeneity):

$$\varphi(\kappa d') \geq \varphi(\kappa d) \quad \text{for all } d' \in N \text{ with } v(\kappa d') = \kappa \lambda.$$

Take  $d'' \in N$  arbitrary with  $v(d'') = \kappa \lambda$  and set  $d' := d''/\kappa \in N$  to deduce

$$\varphi(d'') \geq \varphi(\kappa d) \quad \text{for all } d'' \in N \text{ with } v(d'') = \kappa \lambda.$$

In other words, we have proved

$$\kappa D_\lambda \subset D_{\kappa \lambda} \quad \text{for all } \kappa > 0 \text{ and } \lambda > 0 \tag{1.1.6}$$

(and this holds vacuously if  $D_\lambda = \emptyset$ ). Because  $1/\kappa > 0$ , there also holds

$$\frac{1}{\kappa} D_\lambda \subset D_{\lambda/\kappa} \quad \text{for all } \kappa > 0 \text{ and } \lambda > 0. \tag{1.1.7}$$

The result then follows by taking successively  $\lambda = 1$  in (1.1.6), and  $\lambda = \kappa$  in (1.1.7).  $\square$

**Remark 1.1.6** Note that the result is independent of any convexity assumption. It still holds if the equality constraint is replaced by an inequality  $v(d) \leq \kappa$ . In this case, append a nonnegative slack variable  $r$  and use the trick

$$v(d) \leq \kappa \iff v(d) + r = \kappa \text{ with } r \geq 0.$$

Then the feasible domain is changed to  $(d, r) \in N \times \mathbb{R}^+ =: N'$  (another cone), and we set

$$\varphi'(d, r) \equiv \varphi(r), \quad v'(d, r) := v(d) + r = \kappa. \tag*{$\square$}$$

In other words, the set  $D_1$  of solutions of (1.1.3) is just multiplied by  $\kappa$  if  $\kappa > 0$  replaces 1 in (1.1.3). If  $D_1$  is considered as a set of directions, it can then be considered as *invariant* under this operation.

Note also that, here again, (1.1.3) has at least one optimal solution because it consists of minimizing the continuous  $f'(x, \cdot)$  on the compact unit sphere. With this in mind, we can specify more accurately our algorithmic scheme:

**Algorithm 1.1.7 (Steepest-Descent Scheme)** Start from some  $x_1 \in \mathbb{R}^n$ . Set  $k = 1$ .

STEP 1 (stopping criterion). If  $0 \in \partial f(x_k)$  stop.

STEP 2 (direction-finding). For some norm  $\|\cdot\|$  take  $d_k$  solving (1.1.3) or (1.1.4).

STEP 3 (line-search). Find a stepsize  $t_k > 0$  and a new iterate  $x_{k+1} = x_k + t_k d_k$  such that  $f(x_k + t_k d_k) < f(x_k)$ .

STEP 4 (loop). Replace  $k$  by  $k + 1$  and loop to Step 1.  $\square$

Clearly enough, a stop in Step 1 means that  $x_k$  is optimal. At Step 2,  $d_k$  is a descent direction indeed, because  $\partial f(x_k)$  and  $\{0\}$  are separated, precisely by this  $d_k \neq 0$ . Note also that  $\|\cdot\|$  might be either chosen at each iteration, or fixed a priori before starting the algorithm. We could take, say

$$\|d\|^2 := \langle Q_k d, d \rangle.$$

With  $Q_k := \nabla^2 f(x_k)$  if  $f$  were  $C^2$ , we would obtain Newton's method (remember §II.2.3).

Before proceeding any further, the reader must be warned that the steepest-descent algorithm 1.1.7 is usually *not convergent*, in that the sequence  $\{x_k\}$  need not converge to a minimizer of  $f$ . This will be the subject of §2. Anyway, its behaviour suffers the same deficiencies as any other steepest-descent scheme (remember the end of §II.2.2). In spite of these serious problems, we will study the steepest-descent scheme thoroughly, mainly because it serves as a *basis* for virtually all the minimization methods, certainly for all those in the framework of this book.

Another drawback of Algorithm 1.1.7 is that its implementation requires knowing the full subdifferential. This is even true for any descent scheme, which requires separating the subdifferential from the origin (see Theorems 1.1.2, 1.1.3). Such an operation is trivial when the gradient exists and is available (take the negative gradient!), but becomes a problem by itself if  $\partial f(x)$  is not fully known. For example, observe in Fig. 1.1.1 that not every negative subgradient is a descent direction. This difficulty is actually the key issue; it will be developed in §3, and will motivate Chap. IX.

## 1.2 Solving the Direction-Finding Problem

In this subsection, we consider the question of solving the steepest-descent problem (1.1.3). With its nonlinear and nonconvex constraint, this problem looks rather impractical, even though it must be solved at each execution of Step 2 in Algorithm 1.1.7. Thus, we place ourselves at one given iteration, and we can drop the iteration index  $k$ , which is fixed. In fact we simply ask how to find a steepest-descent direction at a given  $x \in \mathbb{R}^n$ . In order to suggest that  $\partial f(x)$  is just an arbitrary (nonempty) convex compact set, we will use the notation

$$S := \partial f(x);$$

the support function of  $S$  will be indifferently denoted by  $f'(x, \cdot)$  or  $\sigma_S$ .

As an alternative to (1.1.3), consider the nicer, “convexified”, problem

$$\min \{f'(x, d) : \|d\| \leq 1\}$$

or equivalently

$$\min_{\|d\| \leq 1} \max_{s \in S} \langle s, d \rangle. \quad (1.2.1)$$

The next result makes precise the difference between (1.2.1) and (1.1.3).

**Theorem 1.2.1** *The (nonempty) solution-sets of (1.1.3) and (1.2.1) have the following properties:*

- Either the minimal objective-value in (1.2.1) is 0, which means that  $0 \in S$ , or equivalently that  $x$  minimizes  $f$ .
- Or  $0 \notin S$ ; then the solution-sets of (1.1.3) and (1.2.1) coincide (and thus do not contain 0).

PROOF. The first assertion is almost trivial. First note that the minimal objective-value in (1.2.1) is never strictly positive since it cannot be larger than  $f'(x, 0) = 0$ . To say that this minimal value is 0 is to say that  $f'(x, d) \geq 0$  for any  $d$  of norm less than 1, hence for any  $d \in \mathbb{R}^n$  because of positive homogeneity. This in turn is to say that there exists no descent direction, i.e. that  $x$  minimizes  $f$ , which means  $0 \in S = \partial f(x)$ .

Suppose now  $0 \notin S$ , so there exists  $\tilde{d}$  with  $\|\tilde{d}\| \leq 1$  and  $f'(x, \tilde{d}) < 0$ ; by positive homogeneity, we may just assume  $\|\tilde{d}\| = 1$ . Thus, the minimal objective-value is negative in (1.2.1) – as well as in (1.1.3) – and  $d = 0$  cannot be optimal in any of these problems.

Then consider an arbitrary  $d$  with  $f'(x, d) < 0$  and  $\|d\| < 1$ . Set  $d' = d/\|d\|$ . It holds  $\|d'\| = 1$  and

$$f'(x, d') = \frac{f'(x, d)}{\|d\|} < f'(x, d).$$

In other words,  $d'$  is feasible in (1.2.1), and strictly better than  $d$ , which therefore cannot be optimal. This means that (1.2.1) is not changed if its feasible set is restricted to  $\|d\| = 1$ , which is the feasible set in (1.1.3).  $\square$

As a result, suppose that we solve the convexified problem (1.2.1) instead of (1.1.3) or (1.1.4) in Step 3 of the steepest-descent Algorithm 1.1.7. There are two cases:

- If the minimal value in (1.2.1) is 0, then  $x_k$  is optimal and the algorithm can be stopped (note that  $d_k$  may or may not be 0 in this case). The stopping criterion is thus obtained as a by-product, after (1.2.1) is solved; so Step 1 can be incorporated into Step 2.
- Or (1.2.1) gives a negative minimal value, hence a normalized optimal  $d$ , which we are entitled to call  $d_k$  because it is a steepest-descent direction.

Although (1.2.1) is better posed than (1.1.3) (at least it is convex) it is still not so manageable in that it has a nonlinear constraint. Fortunately, the following results shows that (1.2.1) is actually nicer than it appears. We recall the definition of the norm  $\|\cdot\|^*$ , dual to  $\|\cdot\|$  (see §V.3.2):

$$\|s\|^* := \max \{\langle d, s \rangle : \|d\| \leq 1\}.$$

**Theorem 1.2.2** Call  $\hat{S}$  the solution-set of

$$\min \{\|s\|^* : s \in S\} \tag{1.2.2}$$

and take an arbitrary  $\hat{s} \in \hat{S}$ . The solutions of (1.2.1) are those solutions of

$$\min \{\langle \hat{s}, d \rangle : \|d\| \leq 1\} \tag{1.2.3}$$

that lie in the normal cone  $N_S(\hat{s})$  to  $S$  at  $\hat{s}$ .

PROOF. The proof is largely based on §VII.4, see in particular Theorem VII.4.2.5. Consider the set  $\tilde{S} \times D$  of saddle-points of the bilinear function  $(s, d) \mapsto \langle s, d \rangle$ , over the product of compact convex sets  $S$  and  $B = \{d : \|d\| \leq 1\}$ :

$$\begin{aligned} \hat{s} \in \tilde{S} \subset S \quad \text{and} \quad \hat{d} \in D \subset B \quad \text{if and only if} \\ \langle s, \hat{d} \rangle \leq \langle \hat{s}, \hat{d} \rangle \leq \langle \hat{s}, d \rangle \quad \text{for all } s \in S \text{ and } d \in B. \end{aligned} \quad (1.2.4)$$

We know that  $\tilde{S}$  is exactly the solution-set of

$$\begin{aligned} \max_{s \in S} \min_{d \in B} \langle s, d \rangle &\iff \max_{s \in S} \{-\max_{d \in B} \langle -s, d \rangle\} \\ \iff \max_{s \in S} \{-\|s\|^*\} &\iff \max_{s \in S} \{-\|s\|^*\} \end{aligned}$$

which is just (1.2.2); so we conclude that  $\tilde{S} = \hat{S}$ .

We know also that  $D$  is exactly the solution-set of (1.2.1); but from (1.2.4),  $\hat{d} \in D$  if and only if, given  $\hat{s} \in \hat{S}$ , the following two properties hold:

$$\begin{aligned} \langle s, \hat{d} \rangle \leq \langle \hat{s}, \hat{d} \rangle \quad \text{for all } s \in S &\quad [\hat{d} \in N_S(\hat{s})] \\ \langle \hat{s}, \hat{d} \rangle \leq \langle \hat{s}, d \rangle \quad \text{for all } d \in B. &\quad [\hat{d} \text{ solves (1.2.3)}] \quad \square \end{aligned}$$

This result indicates how to solve the convexified steepest-descent problem (1.2.1); now, we show how to recognize a solution.

**Corollary 1.2.3** *The following statements are equivalent:*

- (i)  $\hat{d}$  solves (1.2.1) and  $\hat{s}$  solves (1.2.2);
- (ii)  $\|\hat{d}\| \leq 1$ ,  $\hat{s} \in S$  and there holds

$$\langle \hat{s}, \hat{d} \rangle = -\|s\|^* = \sigma_S(\hat{d}) \quad [= f'(x, \hat{d})]. \quad (1.2.5)$$

PROOF. Theorem 1.2.2 tells us that (i) holds if and only if:

$$\hat{d} \text{ solves (1.2.3), hence } \langle -\hat{s}, \hat{d} \rangle = -\|s\|^* \quad [\text{by definition of the dual norm}]$$

and

$$\hat{d} \in N_S(\hat{s}) \quad \text{i.e.} \quad \langle s, \hat{d} \rangle \leq \langle \hat{s}, \hat{d} \rangle \text{ for all } s \in S.$$

Altogether, this is just (ii).  $\square$

In the language of Proposition V.3.1.4, we have the correspondence  $\hat{s} \in F_S(\hat{d})$ ,  $\hat{d} \in N_S(\hat{s})$ , associating normal cones and exposed faces.

The computation of an  $\hat{s} \in \hat{S}$  in Theorem 1.2.2 is a familiar enough problem: project (in the  $\|\cdot\|^*$ -sense) the origin onto a compact convex set, and obtain a solution  $\hat{s}$ . From Corollary 1.2.3, the solutions of (1.2.1) are then the solutions of the system

$$\left| \begin{array}{l} \|\hat{d}\| \leq 1, \\ \langle \hat{s}, \hat{d} \rangle = -\|\hat{s}\|^*, \\ \langle s, \hat{d} \rangle \leq \langle \hat{s}, \hat{d} \rangle \quad \text{for all } s \in S. \end{array} \right. \quad (1.2.6)$$

The following problem may be considered as more handy than (1.2.6):

$$\begin{cases} \min \|d\|, \\ \langle \hat{s}, d \rangle = -\|\hat{s}\|^*, \\ \langle s, d \rangle \leq \langle \hat{s}, d \rangle \text{ for all } s \in S \quad [\text{i.e. } d \in N_S(\hat{s})], \end{cases} \quad (1.2.7)$$

a convex minimization problem with one affine equality constraint and a possibly infinite number of linear inequality constraints. Once again, this latter problem is “almost” equivalent to the convexified steepest-descent problem (1.2.1):

**Proposition 1.2.4** *Let  $\hat{s}$  solve the projection problem (1.2.2). The solutions of (1.2.7) solve (1.2.1). Conversely, (1.2.1) and (1.2.7) have the same solution-set if  $\hat{s} \neq 0$ .*

PROOF. Because (1.2.1) has a solution, (1.2.6) does have a solution. The optimal value in (1.2.7) is therefore not greater than 1, and any optimal solution of (1.2.7) solves (1.2.1): this is Corollary 1.2.3.

If  $\hat{s} \neq 0$ , all the solutions of (1.2.1) have norm 1 (this is Theorem 1.2.1); hence all the solutions of the equivalent problems (1.2.3) and (1.2.6) have norm 1. We conclude that the minimal value in (1.2.7) is exactly 1 and (1.2.7) is really equivalent to (1.2.6) = (1.2.1).  $\square$

If  $\hat{s} = 0$ , observe that (1.2.7) has the unique solution  $\hat{d} = 0$ . Yet, (1.2.1) or (1.2.6) may have nonzero solutions, unless  $N_S(0) = \{0\}$ , i.e.  $0 \in \text{int } \partial f(x)$ . This confirms that (1.2.7) is not exactly equivalent to (1.2.1). Another observation comes from Proposition 1.1.5: the solutions of (1.2.7) depend multiplicatively on  $\|\hat{s}\|^*$  (the only non-homogeneity in the problem). When  $\hat{s} \neq 0$ , the steepest-descent directions are, up to a normalization, the solutions of

$$\begin{cases} \min \|d\| \\ \langle \hat{s}, d \rangle = -1, \\ \langle s - \hat{s}, d \rangle \leq 0 \text{ for all } s \in S. \end{cases} \quad (1.2.8)$$

Let us sum up this section: to perform Steps 1 and 2 in the steepest-descent Algorithm 1.1.7, one has to

- solve (1.2.2), a projection problem,
- check that it has a nonzero solution  $\hat{s}$  (otherwise stop),
- solve (1.2.6), (1.2.7), or (1.2.8), according to one’s own taste.

**Remark 1.2.5** The constraint  $d \in N_S(\hat{s})$  in (1.2.7) may be really troublesome if  $S$  is complicated enough (but the very possibility of solving (1.2.2) can then be questioned). We mention a situation when it does not bring any trouble, however: suppose the problem

$$\min \{\|d\| : \langle \hat{s}, d \rangle = -1\}$$

has a unique solution  $\hat{d}$ . Then, this  $\hat{d}$  has to lie in  $N_S(\hat{s})$  and solve (1.2.7) (otherwise (1.2.7), hence (1.2.1), would have no solution!). In this case, the last line of constraints in (1.2.7) or (1.2.8) can be neglected.  $\square$

It is convenient to call (1.2.2) the problem *dual* to that of finding a steepest-descent direction, since it is posed in the space of subgradients, the dual space – and as such, it involves the dual norm. As for (1.2.1), or (1.2.8), it is of course the *primal* problem.

### 1.3 Some Particular Cases

The previous section was rather technical, let us see now the practical implications of its results.

**Example 1.3.1** For a first illustration, take in the two-dimensional space  $\mathbb{R}^2$

$$\partial f(x) = \text{co}\{(0, 3/2), (3, 0)\} \quad (1.3.1)$$

and, with  $d = (d_1, d_2) \in \mathbb{R}^2$ ,

$$\|d\| = \|d\|_1 := |d_1| + |d_2|.$$

Assuming that the usual dot-product  $\langle s, d \rangle := s^\top d$  is used, we obtain for  $s = (s_1, s_2)$

$$\|s\|^* = \|s\|_\infty := \max\{s_1, s_2\}.$$

Direct calculations show that (1.2.2) has the unique solution  $\hat{s} = (1, 1)$  (see Fig. 1.3.1, to be compared with Figs. II.2.1.1, II.2.2.1, VII.1.1.2). Then (1.2.3) reads

$$\min_d \{d_1 + d_2 : |d_1| + |d_2| \leq 1\},$$

whose solution-set is the segment  $\text{co}\{-(1, 0), -(0, 1)\}$ . Among these solutions, we have in particular  $-(1/3, 2/3)$ : the  $\ell_1$ -unitary normal to  $\partial f(x)$  at  $(1, 1)$ . According to Theorem 1.2.2, it is the unique steepest-descent direction.

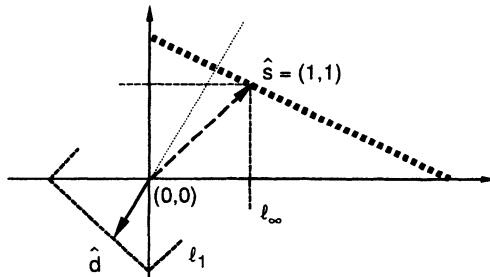


Fig. 1.3.1. The steepest-descent problem with a non-Euclidean norm

Instead of (1.2.3),  $\hat{d}$  could be computed via (1.2.8), which takes the form

$$\begin{cases} \min_d [|d_1| + |d_2|] \\ d_1 + d_2 = -1, \\ -2d_1 + d_2 \leq 0, \\ d_1 - \frac{1}{2}d_2 \leq 0. \end{cases}$$

It is another good exercise to see (graphically or algebraically) that the unique solution of this problem is again  $-(1/3, 2/3)$  – which is  $\ell_1$ -unitary by chance, just because  $\hat{s}$  is  $\ell_\infty$ -unitary.  $\square$

**Example 1.3.2** The above example can be transformed in a pernicious way by taking  $\partial f(x) = \{(1, 1)\}$ . Because  $f$  is differentiable at  $x$ , the situation seems to become simpler; but this is somewhat misleading. In fact, (1.2.2) still has the unique solution  $(1, 1)$ ; the segment  $\text{co}\{-(1, 0), -(0, 1)\}$  is still the solution-set of (1.2.3). Now, however, this whole segment forms the set of steepest-descent directions because it is entirely contained in the normal cone to  $(1, 1)$ , namely the whole of  $\mathbb{R}^2$ . We have a confirmation that uniqueness of a steepest-descent direction has nothing to do with differentiability of  $f$ .

Rather, this uniqueness is related with the selected norm. Take again  $\partial f(x)$  from (1.3.1) but now with the Euclidean norm:  $\|\cdot\| = \|\cdot\|$ . Then it becomes obvious that (1.2.2) has the unique solution  $\hat{s} = (3/5, 6/5)$  and that (1.2.7) has the unique solution  $\hat{d} = -(1/3, 2/3)$ .

It is interesting to observe that this last steepest-descent direction, obtained with the  $\ell_2$  norm, is the same as the first one, obtained with the  $\ell_1$  norm. Yet, the dual solution  $\hat{s}$  of (1.2.2) does depend on the norming. The real explanation of this paradox is that we are in  $\mathbb{R}^2$ : for “most”  $\hat{s} \in \partial f(x)$ , the normal cone  $N_{\partial f(x)}(\hat{s})$  is the same straight line  $(1, 2)\mathbb{R}$ .  $\square$

**Example 1.3.3** Take now a case without uniqueness in the dual problem (1.2.2):  $\partial f(x)$  being still (1.3.1), let the norming be

$$\|s\|^* := |s_1| + 2|s_2|$$

which is the dual of

$$\|d\| := \max \left\{ |d_1|, \frac{1}{2}|d_2| \right\}.$$

In this case, all the subgradients have the same dual norm. If, when solving (1.2.2), we obtain  $\hat{s} \in \text{ri } \partial f(x)$  – for example  $\hat{s} = (1, 1)$  – then (1.2.7) has the unique solution  $\hat{d} = -(1/3, 2/3)$  as before.

Suppose, on the other hand, that it is  $\hat{s} = (3, 0)$  that crops up from (1.2.2). Then  $N_S(\hat{s})$  becomes a half-space and uniqueness in the primal problem is less obvious. In this case, (1.2.3) reads

$$\begin{cases} \min_d 3d_1 \\ \max \left\{ |d_1|, \frac{1}{2}|d_2| \right\} \leq 1, \end{cases}$$

whose solution-set is the whole segment  $[-(1, -2), (-1, 2)]$ . Its intersection with  $N_S(3, 0)$  is, as by chance, the singleton  $-(1, 2)$ ! Likewise, (1.2.7) is

$$\begin{cases} \min_d \max \left\{ |d_1|, \frac{1}{2}|d_2| \right\} \\ 3d_1 = -3, \\ 3d_1 - \frac{3}{2}d_2 \geq 0, \end{cases}$$

which, another “chance”, has the unique solution  $-(1/3, 2/3)$ ! We leave it to the reader to draw the appropriate pictures and to work out the complete calculations.  $\square$

These examples show that the concept of steepest-descent direction is more intuitive when the norm used is the Euclidean norm  $\|\cdot\| = \langle \cdot, \cdot \rangle^{1/2}$ : it yields uniqueness, and it suggests pictures agreeing with classical geometry. Closely related is the important case of a general quadratic normalization: take

$$\|d\|^2 := \langle Qd, d \rangle \quad (1.3.2)$$

where  $Q$  is a symmetric positive definite linear operator. Then (see Example V.3.2.3), for  $s \neq 0$ , the problem

$$\max \{ \langle s, d \rangle : \langle Qd, d \rangle \leq 1 \}$$

has the unique solution  $d(s) = \frac{Q^{-1}s}{\langle s, Q^{-1}s \rangle^{1/2}}$  and the dual norm of  $\|\cdot\|$  is

$$\|s\|^* = \sqrt{\langle s, Q^{-1}s \rangle}.$$

In this framework, the results of §1.2 can be copied. As a by-product, we obtain first the following classical projection theorem.

**Proposition 1.3.4** *Let  $S$  be a nonempty compact convex set and  $Q$  a symmetric positive definite linear operator. There is only one  $s \in S$  satisfying*

$$\sigma_S(-Q^{-1}s) = -\langle s, Q^{-1}s \rangle \quad (1.3.3)$$

and it is the unique solution of

$$\min \{ \langle s, Q^{-1}s \rangle : s \in S \}. \quad (1.3.4)$$

PROOF. The constrained minimization problem (1.3.4) has a unique solution  $\hat{s}$  (the objective function is strictly convex). From Theorem VII.1.1.1,  $\hat{s}$  is characterized by:  $\hat{s} \in S$  and

$$\langle -Q^{-1}\hat{s}, s - \hat{s} \rangle \leq 0 \quad \text{for all } s \in S.$$

This is exactly (1.3.3).  $\square$

Then characterizing the solutions of (1.2.1) becomes trivial.

**Proposition 1.3.5** *Let  $S$  be a nonempty convex compact set and  $Q$  a symmetric positive definite linear operator. The solution-set  $D$  of*

$$\min \{ \sigma_S(d) : \langle Qd, d \rangle \leq 1 \}$$

is characterized as follows.

(i) If  $0 \notin S$ ,  $D$  reduces to the unique  $Q$ -normalized vector along  $-Q^{-1}\hat{s}$ , with  $\hat{s} \neq 0$  solving (1.3.4); say

$$\hat{d} = \frac{-Q^{-1}\hat{s}}{\sqrt{\langle \hat{s}, Q^{-1}\hat{s} \rangle}}.$$

(ii) If  $0 \in S$ ,  $D$  is the truncated normal cone to  $S$  at 0, i.e. the set of  $d$  with  $Q$ -norm not exceeding 1, for which  $\sigma_S(d) = 0$ :

$$D = \{d \in N_S(0) : \langle Qd, d \rangle \leq 1\}.$$

In particular, if  $0 \in \text{int } S$ , then  $D = \{0\}$ .  $\square$

**Remark 1.3.6** We indicate a practical mnemonic to find quickly if it is  $Q$  or  $Q^{-1}$  that is involved in the above formulae: directions  $d$  are in the space  $\mathbb{R}^n$  considered as primal, while subgradients  $s$  are in its dual;  $Q$  in (1.3.2) sends a primal vector into the dual space,  $Q^{-1}$  sends a dual vector back into the primal space. In this interpretation, applying  $Q$  to  $s$ , say, would have no meaning.

If we assume that  $\nabla f(x)$  and  $\nabla^2 f(x)$  exist and if we take  $Q = \nabla^2 f(x)$ , then we obtain  $\hat{d}$  collinear to  $-[\nabla^2 f(x)]^{-1} \nabla f(x)$ , the Newton direction.  $\square$

The purely Euclidean case  $\|\cdot\| = \langle \cdot, \cdot \rangle^{1/2}$  is of course obtained with  $Q = I_n$ , the identity operator. Switching to the notation  $\partial f(x)$  and  $f'(x, d)$ , the above formulae reduce to

$$\hat{s} = \underset{s \in \partial f(x)}{\operatorname{argmin}} \frac{1}{2} \|s\|^2; \quad \hat{d} = \frac{-\hat{s}}{\|\hat{s}\|}; \quad f'(x, \hat{d}) = \langle \hat{s}, \hat{d} \rangle = -\|\hat{s}\|.$$

In this case,  $\hat{s}$  can be interpreted as playing the role of “the gradient”. Suppose  $\hat{s} \neq 0$  and let  $d$  be an arbitrary direction of norm 1. Because  $\hat{d}$  is the steepest-descent direction, there holds

$$f'(x, d) \geq f'(x, \hat{d}) = -\|\hat{s}\|$$

with equality if  $d = \hat{d}$ . This displays an interpretation of the “norm of the gradient”  $\|\hat{s}\|$ : it is the fastest possible rate of decrease of  $f$  along normalized directions issued from  $x$ .

**Remark 1.3.7** Let us comment this point: suppose that  $f$  has a gradient at  $x$  and set  $\hat{s} := \nabla f(x)$ . Then for any normalized direction  $d$ , there holds

$$f'(x, d) = \langle \hat{s}, d \rangle \geq -\|\hat{s}\| \|d\| = -\|\hat{s}\|. \quad (1.3.5)$$

Of course,  $\|\hat{s}\|$  is still the maximal decrease of  $f$  around  $x$ , but note that the bound in (1.3.5) is solely due to the Cauchy-Schwarz inequality. The “loss in optimality” of a direction is directly driven by its angle with the optimal  $\hat{d}$ .

Suppose, on the contrary, that  $\partial f(x)$  is not a singleton and reconstruct the calculation (1.3.5) above (see again Fig. 1.1.2). Assuming that the maximum  $f'(x, d)$  of  $\langle \cdot, d \rangle$  over  $\partial f(x)$  is attained at some  $s(d)$ , we obtain

$$f'(x, d) = \langle s(d), d \rangle \geq \langle \hat{s}, d \rangle \geq -\|\hat{s}\| \|d\| = -\|\hat{s}\|.$$

An additional inequality has come into play, to account for the fact that  $s(d)$  may be different from  $\hat{s}$ . It “worsens”  $d$ , seen as a descent direction;  $f'(x, d)$  may even become positive for  $d$  rather close to  $\hat{d}$  (see Fig. 1.3.2) – one more nasty feature of non-differentiability.  $\square$

**Remark 1.3.8** Case (ii) of Proposition 1.3.5 suggests the following observation. As already mentioned in Remark VI.2.2.5, an important object associated with an optimal point is the normal cone  $N_S(0)$ : it is the cone of critical directions, along which  $f$  looks constant.

When solving (1.2.1) instead of (1.1.3), we may get  $\hat{s} = 0$ : then we learn that the current  $x$  is optimal. The mere knowledge of  $\hat{s} = 0$ , however, means a loss of information concerning a *parametric* study of the optimal  $x$ . The place of  $\hat{s} = 0$  in  $S$  may not be known; for example, the important property: “is 0 in  $\text{int } S$ ?” may not be answered. We may not know the set  $N_S(0)$  of “dangerous” directions either.

On the other hand, an optimal solution  $\tilde{d}$  of (1.1.3) is never zero. If its optimal value  $f'(x, \tilde{d})$  is positive, we obtain the minimal rate of *increase* of  $f$  around  $x$ , i.e. the worst  $\varepsilon$  in (VI.2.2.4) of Remark VI.2.2.4. If  $f'(x, \tilde{d}) = 0$ , this  $\tilde{d}$  is a critical direction.  $\square$

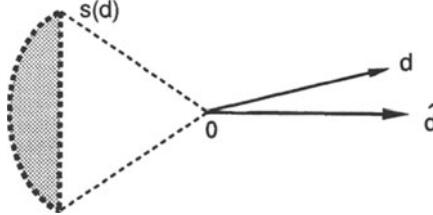


Fig. 1.3.2. Discontinuity of an exposed face

## 1.4 Conclusion

We have shown in this Section 1 that computing the direction in the steepest-descent Algorithm 1.1.7 amounts to solving two optimization problems: first the projection (1.2.2) and then (1.2.8). It is now necessary to ask the question: is this a *constructive* way of computing a steepest-descent direction?

Both problems (1.2.2) and (1.2.8) involve the structure of the norming *and* of the subdifferential. However, (1.2.8) can be considered as the easier problem, because its complexity depends less on  $\partial f(x)$ . For example, suppose we solve (1.2.3), which does not involve the subdifferential, and obtain a unique solution – as is the case with a quadratic norming. Then this solution is the required steepest-descent direction (remember Remark 1.2.5). Finally note that, without a strong motivation for the contrary, the norm should simply be the Euclidean one: the analysis in §II.2.2(c) suggests that any other “off line” norm should perform more poorly.

On the other hand, (1.2.2) is usually impossible to solve, unless there is some structure in  $f$  (which is not under our control!). We mention three instances in which such a manageable structure exists. For the sake of simplicity, we assume the norming to be  $\|\cdot\| \equiv \|\cdot\| = \langle \cdot, \cdot \rangle^{1/2}$ .

*Case 1.* The subdifferential is a compact convex polyhedron characterized as a convex hull:  $s_1, s_2, \dots, s_m$  are given in  $\mathbb{R}^n$  and,  $\Delta_m$  being the unit simplex,

$$\partial f(x) := \text{co}\{s_1, \dots, s_m\} = \left\{ s = \sum_{j=1}^m \alpha_j s_j : \alpha \in \Delta_m \right\}.$$

Then (1.2.2) is the convex quadratic minimization problem with  $m$  variables  $\alpha_j$

$$\min_{\alpha \in \Delta_m} \frac{1}{2} \left\| \sum_{j=1}^m \alpha_j s_j \right\|^2.$$

*Case 2.* The subdifferential is a convex polyhedron (assumed compact) characterized by its supporting hyperplanes:  $m$  nonzero vectors  $v_1, v_2, \dots, v_m$  are given in  $\mathbb{R}^n$ , together with  $m$  numbers  $r_1, r_2, \dots, r_m$  and

$$\partial f(x) := \{s : \langle s, v_j \rangle \leq r_j \text{ for } j = 1, 2, \dots, m\}.$$

Then (1.2.2) is again a convex quadratic minimization problem, but this time with  $n$  variables and  $m$  inequality constraints

$$\min \left\{ \frac{1}{2} \|s\|^2 : \langle s, v_j \rangle \leq r_j \text{ for } j = 1, 2, \dots, m \right\}.$$

*Case 3.* The subdifferential is an ellipsoid:

$$\partial f(x) := \{s = Rz + c : \|z\| \leq 1\}$$

where  $R : \mathbb{R}^m \rightarrow \mathbb{R}^n$  is a given linear mapping. Then (1.2.2) reads

$$\min \left\{ \frac{1}{2} \|Rz + c\|^2 : \frac{1}{2} \|z\|^2 \leq 1/2 \right\}. \quad (1.4.1)$$

It is a good exercise to study this problem. Taking a Lagrange multiplier  $\mu$ , set  $Q = Q(\mu) := R^*R + \mu I_n$  and write the minimality conditions

$$Qz + R^*c = 0, \quad \|z\| \leq 1, \quad \mu \geq 0, \quad \mu = 0 \text{ if } \|z\| < 1.$$

Because  $R^*R$  is symmetric positive semi-definite,  $Q(\mu)$  is invertible for all  $\mu > 0$ , but  $Q(0)$  may be singular.

Let  $z_0$  be the solution of the convex quadratic minimization problem

$$\min \left\{ \frac{1}{2} \|z\|^2 : Q(0)z + R^*c = 0 \right\}.$$

- If  $\|z_0\| \leq 1$ , then  $z_0$  solves (1.4.1) and our requested projection is  $\hat{s} = Rz_0 + c$ .
- If  $\|z_0\| > 1$ , no  $z$  with norm less than 1 can solve the minimality conditions: we must have  $\mu > 0$ , so the nonlinear equation

$$\|z\| = \|[Q(\mu)]^{-1}R^*c\| = 1$$

has a solution  $\hat{\mu} > 0$ . Then we obtain  $\hat{s} = -R[Q(\hat{\mu})]^{-1}R^*c + c$ .

In this second situation, finding the exact  $\hat{\mu}$  is of course not possible but approximating it is an easy task via, for example, a univariate Newton method.

Let us conclude this section:

- a steepest-descent algorithm is associated with a particular norm;
- it is a non-convergent algorithm (see §2 below);
- to implement it, one needs to characterize the full subdifferential (see §3 below);
- the direction is computed essentially by projecting the origin onto this subdifferential, in the sense of the dual of the norm considered;
- this can be conveniently done only when the subdifferential is a closed convex polyhedron, or also an ellipsoid.

The next section, precisely, will study the case when the subdifferential is a computable polyhedron.

## 2 Illustration. The Finite Minimax Problem

In this section, we address the problem of minimizing a function having the special form

$$f(x) := \max \{f_j(x) : j = 1, \dots, p\} \quad (2.0.1)$$

where each  $f_j : \mathbb{R}^n \rightarrow \mathbb{R}$  is convex and (continuously) differentiable, and  $p$  is some given positive number. To minimize such an  $f$  is what we call a *finite minimax problem* (whereas, in a general minimax problem,  $j$  could range over an infinite set).

We assume throughout that all the functions  $f_j$  are available, together with their gradients. In the terminology of Chap. II (and more particularly Fig. II.1.2.1), this means that the black box (U1) is much more elaborate than usual: instead of one number  $f(x)$  and one vector  $s(x)$ , it computes  $p$  numbers  $f_j(x)$  and  $p$  vectors  $\nabla f_j(x)$ .

## 2.1 The Steepest-Descent Method for Finite Minimax Problems

For each  $x \in \mathbb{R}^n$ , we denote by

$$J(x) := \{j : f_j(x) = f(x)\} \quad (2.1.1)$$

the *active index-set* at  $x$ . The functions  $f_j$  and gradients  $\nabla f_j$  for  $j \in J(x)$  will be called respectively the active functions and active gradients (at  $x$ ).

The fundamental Corollary VI.4.3.2 then gives:

**Theorem 2.1.1** *The function  $f$  of (2.0.1) is convex. For given  $x$ , its subdifferential is the convex hull of the active gradients at  $x$ :*

$$\partial f(x) = \text{co} \{ \nabla f_j(x) : j \in J(x) \}.$$

□

Thus, the subdifferential of such an  $f$  is a compact convex polyhedron, having at most  $p$  extreme points. An actual computation of this polyhedron exactly amounts to performing the following operations, which can be done by a computer program:

- find all the active indices  $j$  at the given  $x$ ;
- do some ordinary differential calculus to compute the corresponding gradients;
- the subdifferential is then the set of all convex combinations of these gradients:

$$\partial f(x) = \left\{ \sum_{j \in J(x)} \alpha_j \nabla f_j(x) : \sum_{j \in J(x)} \alpha_j = 1, \alpha_j \geq 0 \text{ for } j \in J(x) \right\}. \quad (2.1.2)$$

It is now purely mechanical to obtain the minimality conditions from Theorem 2.1.1:

**Proposition 2.1.2** *A necessary and sufficient condition for  $x$  to minimize  $f$  defined by (2.0.1) is that there exist coefficients  $\alpha_j$ ,  $j \in J(x)$ , satisfying:*

$$\alpha_j \geq 0 \text{ for } j \in J(x), \quad \sum_{j \in J(x)} \alpha_j = 1, \quad \sum_{j \in J(x)} \alpha_j \nabla f_j(x) = 0.$$

PROOF. Trivial from (2.1.2):  $x$  minimizes  $f$  if and only if  $0 \in \partial f(x)$ . □

**Remark 2.1.3** As already observed in §VI.2.1,  $\partial f(x)$  is a singleton whenever  $J(x)$  is a singleton, say  $j(x)$ ; then  $\partial f(x) = \{\nabla f(x)\} = \{\nabla f_{j(x)}(x)\}$ . This may serve as an intuitive explanation of Theorem IV.4.2.3, stating that a convex function is differentiable almost everywhere: equality between two real numbers (here:  $f_i(x)$  and  $f_j(x)$  for some  $i \neq j$ ) is an “extraordinary” event.

It would be dangerous, however, to minimize  $f$  of (2.0.1) by means of mere “smooth tools”, under the pretext that “it should work almost surely”. Remark VI.2.2.2 has mentioned that, as a rule, a convex function is not differentiable at a minimum point. Such is the case, for example, if each  $f_j = \langle s_j, \cdot \rangle + r_j$  of (2.0.1) is affine but not constant: then a differentiability point cannot be a minimum, since the gradient of  $f$  has to be some  $s_j \neq 0$ . In words, a “smooth tool” is almost never appropriate in nonsmooth optimization.

Another consequence of the above remark concerns the algorithmic scheme of §1. It has been already mentioned that steepest descent may not converge to an optimal point. Here is a first explanation: assuming the Euclidean norming, the sequence  $\{x_k\}$  is generated by

$$x_{k+1} = x_k - t_k s_k;$$

$s_k$  is some subgradient of  $f$  at  $x_k$  (having the shortest norm) and  $t_k$  is given by some line-search, say inspired from §II.3. Now, Theorem IV.4.2.3 tells us that  $s_k$  is “likely” to be simply the gradient  $\nabla f(x_k)$ . Thus, *in practice*, we have gained nothing when defining the steepest-descent Algorithm 1.1.7: it trivially reduces to the gradient method of Definition II.2.2.2 and becomes highly suspect!  $\square$

The next direct consequence of Theorem 2.1.1 is that a descent direction is relatively easy to construct:

**Proposition 2.1.4** *A descent direction is a  $d$  satisfying the finite set of inequalities*

$$\langle d, \nabla f_j(x) \rangle < 0 \quad \text{for all } j \in J(x). \quad (2.1.3)$$

*Actually, there holds*

$$f'(x, d) = \max \{ \langle \nabla f_j(x), d \rangle : j \in J(x) \}. \quad (2.1.4)$$

**PROOF.** This result is fairly trivial from §VI.1: the linear function  $\langle d, \cdot \rangle$  attains its maximum  $f'(x, d)$  over some extreme point of the convex polyhedron (2.1.2). Nevertheless, we sketch an elementary proof of (2.1.4) to explain this important result once again: the following proof can be read without knowing anything from Chap. VI.

In what follows,  $t > 0$  is small enough. By continuity of each  $f_j$ , those indices not in  $J(x)$  do not count at  $x + td$ , i.e.

$$f(x + td) > f_j(x + td) \quad \text{for all } j \notin J(x).$$

This means that  $J(x + td) \subset J(x)$ , so we can replace (2.0.1) by

$$f(x + td) = \max \{ f_j(x + td) : j \in J(x) \}, \quad (2.1.5)$$

valid around  $x$ . Now a first-order development of  $f_j$  yields

$$\begin{aligned} f(x + td) &= \max \{f_j(x) + t\langle \nabla f_j(x), d \rangle + t\varepsilon_j(t) : j \in J(x)\} \\ &= f(x) + t \max \{\langle \nabla f_j(x), d \rangle + \varepsilon_j(t) : j \in J(x)\} \end{aligned}$$

where  $\varepsilon_j(t) \rightarrow 0$  for  $t \downarrow 0$ . Letting  $t \downarrow 0$  proves (2.1.4). As for the descent property (2.1.3), it is a consequence of (2.1.4).  $\square$

Of course, the above “proof” hardly generalizes to the case of an infinite set of indices. Then, there are two difficult points: (2.1.5) must be replaced by an asymptotic property, namely all the indices in  $J(x + td)$  cluster to  $J(x)$  when  $t \downarrow 0$ ; and the convergence of each  $\varepsilon_j(\cdot)$  towards 0 must be uniform in  $j$ .

We are now in a position to specify the algorithmic scheme of §1 in some detail.

**Algorithm 2.1.5 (Steepest-Descent, Finite Minimax Problems)** The initial point  $x_1 \in \mathbb{R}^n$  and the tolerance  $\delta > 0$  are given, together with the black box (U1) which, given  $x$ , computes  $f(x)$ ,  $J(x)$  and the gradients  $\nabla f_j(x)$ ,  $j \in J(x)$ . Set  $k = 1$ .

STEP 1 (projection and stopping criterion). For some norm  $\|\cdot\|$  solve the projection problem

$$\left| \begin{array}{l} \min \|\sum_{j \in J(x_k)} \alpha_j \nabla f_j(x_k)\|^* \\ \sum_{j \in J(x_k)} \alpha_j = 1, \\ \alpha_j \geq 0 \text{ for } j \in J(x_k), \end{array} \right. \quad (2.1.6)$$

and call  $s_k := \sum_{j \in J(x_k)} \alpha_j \nabla f_j(x_k)$  the result. If  $\|s_k\|^* \leq \delta$  stop.

STEP 2 (direction finding). Solve

$$\left| \begin{array}{l} \min \|d\| \\ \langle s_k, d \rangle = -1, \\ \langle \nabla f_j(x_k), d \rangle \leq -1 \text{ for } j \in J(x_k) \end{array} \right. \quad (2.1.7)$$

with respect to  $d$ , and obtain a solution  $d_k \neq 0$ .

STEP 3 (line-search). Find a stepsize  $t_k > 0$  and a new iterate  $x_{k+1} = x_k + t_k d_k$  such that  $f(x_k + t_k d_k) < f(x_k)$ .

STEP 4 (loop). Replace  $k$  by  $k + 1$  and loop to Step 1.  $\square$

Of course, Step 2 is a disguised form of the convexified steepest-descent problem (1.2.8): the inequality-constraints of the  $d$ -problem are really

$$\langle \nabla f_j(x_k) - s_k, d \rangle \leq 0 \quad \text{for } j \in J(x_k),$$

i.e.  $d \in N_{\partial f(x_k)}(s_k)$ .

Let us recall again that the above algorithm is not convergent; we skip therefore the problem of finding  $t_k$  in Step 3. It is a detail to be reserved for subsequent chapters, when we study more reasonable (but closely related) algorithms for nonsmooth optimization. For the moment, it suffices to say that a line-search can be implemented according to the principles of §II.3.

**Remark 2.1.6** Our primary problem was to minimize the function  $x \mapsto f(x)$  of (2.0.1) over the primal space  $\mathbb{R}^n$ . After linearization around some given  $x = x_k$ , it became that of minimizing the function  $d \mapsto f'(x, d)$ , and this was the (primal) minimax problem coming from (1.2.1):

$$\min_{\|d\| \leq 1} \max_{\alpha \in \Delta} \left( \sum_{j \in J(x)} \alpha_j \nabla f_j(x), d \right). \quad (2.1.8)$$

Here  $\Delta$  denotes the set of convex multipliers indexed in  $J(x)$ .

Dually, we have the associated maximin problem coming from (1.2.2):

$$\max_{\alpha \in \Delta} \min_{\|d\| \leq 1} \left( \sum_{j \in J(x)} \alpha_j \nabla f_j(x), d \right),$$

which is just (2.1.6) (barring the sign). Needless to say, our primal-dual terminology comes from the theory of saddle-points in §VII.4:  $d$  and  $\alpha$  are primal and dual variables respectively,  $\langle \sum_{j \in J(x)} \alpha_j s_j, d \rangle$  being a Lagrangian. For  $j \in J(x)$ ,  $\alpha_j$  is the Lagrange multiplier associated with the  $j^{\text{th}}$  linear constraint in the primal problem (2.1.8) = (1.2.1):

$$\begin{cases} \min r & (d, r) \in \mathbb{R}^n \times \mathbb{R}, \\ \langle \nabla f_j(x), d \rangle \leq r & \text{for } j \in J(x), \\ \|d\| \leq 1. \end{cases} \quad (2.1.9)$$

Indeed, the  $r$ -part of the minimality conditions readily gives  $\alpha \in \Delta$ .

For a general convex function  $f$ ,  $\partial f(x)$  is no longer a compact convex polyhedron; the linear constraints of (2.1.9) become

$$\langle s, d \rangle \leq r \quad \text{for all } s \in \partial f(x),$$

with an infinite “index set”  $\partial f(x)$ . It is now harder to speak of Lagrange multipliers. Nevertheless, the message coming from Theorems VII.4.3.1 and VII.4.5.1 is that a Lagrangian and a dual problem can still be defined. The trick is that the would-be  $\alpha \in \Delta$  gives birth to  $s = \sum_j \alpha_j s_j$ , describing all convex combinations of points  $s_j \in \partial f(x)$ . It is now  $s$  that becomes the set of multipliers.

A last remark, in anticipation to subsequent chapters: after all, the saddle-point theory of §VII.4 is not restricted to functions  $\ell$  which are affine with respect to one argument. As far as saddle-points are concerned, affinity of the Lagrangian  $L(x, \cdot)$  is not compulsory. The present dualization business could therefore be done *before* linearization: it could be applied directly to the problem  $\min f(x)$ , instead of the problem of minimizing the function  $d \mapsto f'(x, d)$ . For this, it should suffice to express  $f$  as a max-function, say

$$\min f(x) \iff \min_x \max_y h(x, y);$$

the dualization would then consist in interchanging the min and max operations.  $\square$

When  $\|\cdot\|$  is chosen as the Euclidean norm  $\|\cdot\| = \langle \cdot, \cdot \rangle^{1/2}$ , the calculations can be worked out more completely. At a given  $x$  (assumed non-optimal), we know from Corollary 1.3.4 the steepest-descent direction  $\hat{d}$ : it is unique and opposite to the Euclidean projection  $\hat{s}$  of the origin onto  $\partial f(x)$ . Consider the problem ( $\Delta$  being the unit simplex introduced in Remark 2.1.6)

$$\min_{\alpha \in \Delta} \frac{1}{2} \left\| \sum_{j \in J(x)} \alpha_j \nabla f_j(x) \right\|^2. \quad (2.1.10)$$

It may have several solutions, but they all make up the same vector:

$$\sum_{j \in J(x)} \hat{\alpha}_j \nabla f_j(x) = \hat{s} \quad \text{for any solution } \hat{\alpha} \text{ of (2.1.10),}$$

which is of course the Euclidean projection of 0 onto the convex hull of the active gradients at  $x$ . Note the following characterization of  $\hat{s}$ :

**Proposition 2.1.7** *The above projection  $\hat{s}$  is the unique convex combination  $s \in \text{co}\{\nabla f_j(x) : j \in J(x)\}$  satisfying*

$$\langle s, \nabla f_j(x) \rangle \geq \|s\|^2 \quad \text{for all } j \in J(x). \quad (2.1.11)$$

*Equality holds in (2.1.11) for all  $j$  such that there is some  $\hat{\alpha}$  solving (2.1.10) and having  $\hat{\alpha}_j > 0$ .*

PROOF. Just apply Proposition 1.3.4 with  $Q = I$  to the convex hull  $S$  of the active gradients. Observe that the inequality

$$\langle s', d \rangle \geq r \quad \text{for all } s' \in S$$

is equivalent to

$$\langle \nabla f_j(x), d \rangle \geq r \quad \text{for all } j \in J(x).$$

This establishes (2.1.11). Now take  $\hat{\alpha}$  solving (2.1.10); if we had some  $j_0$  with  $\hat{\alpha}_{j_0} > 0$  and strict inequality holding in (2.1.11), then we would have for this  $j_0$

$$\langle \hat{s}, \hat{\alpha}_{j_0} \nabla f_{j_0}(x) \rangle > \hat{\alpha}_{j_0} \|\hat{s}\|^2.$$

Multiplying the other inequalities by  $\hat{\alpha}_j$  and summing, we would obtain the contradiction

$$\langle \hat{s}, \hat{s} \rangle > \|\hat{s}\|^2. \quad \square$$

Call  $J_1(x)$  the set of indices alluded to in Proposition 2.1.7:

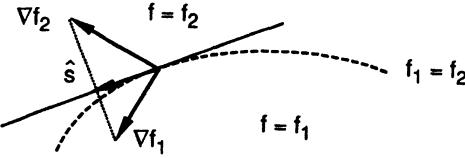
$$J_1(x) := \{j \in J(x) : \exists \hat{\alpha} \text{ solving (2.1.10) with } \hat{\alpha}_j > 0\}.$$

All the gradients  $\nabla f_j(x)$  with  $j \in J_1(x)$  lie in the face of  $\partial f(x)$  exposed by  $\hat{s}$ . The property that  $\langle \nabla f_j(x), \hat{s} \rangle$  is independent of  $j \in J_1(x)$  can be written

$$f_j(x) + \langle \nabla f_j(x), \hat{s} \rangle \text{ is constant for any } j \in J_1(x).$$

Its geometric meaning (see Fig. 2.1.1) is that  $\hat{s}$  is tangent to the surface defined by

$$\{h \in \mathbb{R}^n : f_j(x + h) = f_i(x + h) \text{ for any } i \text{ and } j \text{ in } J_1(x)\}.$$



**Fig. 2.1.1.** Steepest descent is tangent to the kinky surface

**Remark 2.1.8** If  $J_1(x)$  were known a priori,  $\hat{s}$  could be computed as the projection of the origin onto the *affine hull* of the corresponding gradients. Instead of (2.1.10), one could solve a problem with variables indexed in  $J_1(x)$  only:

$$\min \left\{ \frac{1}{2} \left\| \sum_{j \in J_1(x)} \alpha_j \nabla f_j(x) \right\|^2 : \sum_{j \in J_1(x)} \alpha_j = 1 \right\}, \quad (2.1.12)$$

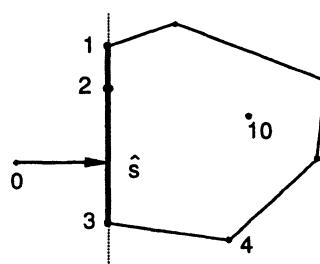
and set to 0 all the other  $\alpha_j$  for  $j \in J \setminus J_1$ . Then  $\hat{s}$  would be equal to  $\sum_{J_1(x)} \alpha_j \nabla f_j(x)$ . The reason appears in Fig. 2.1.2, in which  $J_1(x) = \{1, 2, 3\}$ : if  $J_1(x)$  has really been correctly chosen, (2.1.12) gives a point in the convex hull

$$\text{co} \{ \nabla f_j(x) : j \in J_1(x) \} \subset \text{co} \{ \nabla f_j(x) : j \in J(x) \}.$$

Note that (2.1.12) does not involve any inequality constraint. It can therefore be solved as a linear system of equations in  $\alpha$ , say

$$\sum_{j \in J_1(x)} \alpha_j \langle \nabla f_i(x), \nabla f_j(x) \rangle = \lambda \quad \text{for all } i \in J_1(x) \quad (2.1.13)$$

where  $\lambda$  (the Lagrange multiplier, which is a posteriori  $\|\hat{s}\|^2$ ) is adjusted so that the  $\alpha_j$ 's sum up to 1. Then (2.1.11) is automatically satisfied for  $j \in J_1(x)$ . If some other constraint in (2.1.11) appears to be violated,  $J_1(x)$  was wrong; so would be the case in Fig. 2.1.2 if we had taken  $J_1(x) = \{1, 4\}$ . Likewise, if no solution of (2.1.13) is nonnegative (we do not assume here that (2.1.13) has a unique solution), this indicates that  $J_1(x)$  was wrong as well; take  $J_1(x) = \{1, 2\}$  in Fig. 2.1.2.  $\square$



**Fig. 2.1.2.** Guessing the correct exposed face

## 2.2 Non-Convergence of the Steepest-Descent Method

We now study a counter-example to support the claim, made on several occasions, that the steepest-descent scheme may not be convergent. Our counter-example shows that such is the case even if the function is as simple as piecewise affine.

Consider the following five functions of  $x = (\xi, \eta) \in \mathbb{R}^2$ :

$$f_0(x) := -100; \quad f_{\pm 1}(x) := \pm 2\xi + 3\eta; \quad f_{\pm 2}(x) := \pm 5\xi + 2\eta \quad (2.2.1)$$

and set, as in (2.0.1)

$$f(x) := \max \{f_0(x), f_{-1}(x), f_{-2}(x), f_1(x), f_2(x)\}.$$

Let us concentrate on the region  $\eta \geq 0$ , in which  $f$  is nonnegative and  $f_0$  does not count. There,  $\nabla f$  fails to exist on the three half-lines

$$L_{\pm} := \{x : 0 \leq \eta = \pm 3\xi\}, \quad \text{where } f_{\pm 1} = f_{\pm 2}$$

and

$$L_0 := \{x : \xi = 0\}, \quad \text{where } f_{-1} = f_1.$$

This is illustrated by Fig. 2.2.1, which shows a level-set of  $f$ , the three critical lines  $L_0$  and  $L_{\pm}$ , and the four possible values for  $\nabla f$ . In the region  $\eta \leq 0$ ,  $f_{\pm 1}$  do not count and  $L_0$  becomes the only critical line;  $L_-$  and  $L_+$  coalesce at 0. Finally, the minimal value of  $f$  is clearly  $-100$ , attained for sufficiently negative values of  $\eta$ .

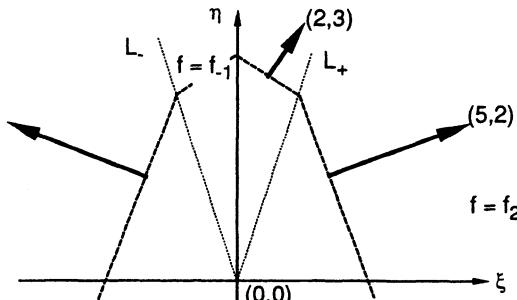
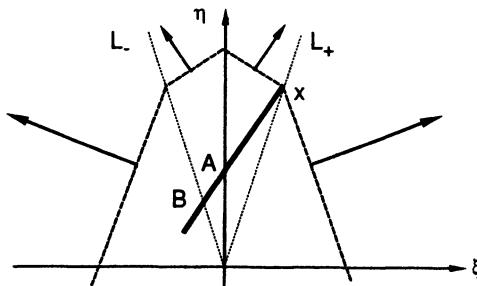


Fig. 2.2.1. A counter-example for steepest descent

Let the current point  $x$  of Algorithm 2.1.5 be in the first quadrant and such that  $f_1$  is active: suppose for example that  $x \in L_+$ . Then the steepest-descent direction is  $-(2, 3)$ ; and it is important to observe that it is still  $-(2, 3)$  even if  $x \notin L_+$ . This is due to the property

$$16 = \langle \nabla f_1(x), \nabla f_2(x) \rangle > \|\nabla f_1(x)\|^2 = 13,$$

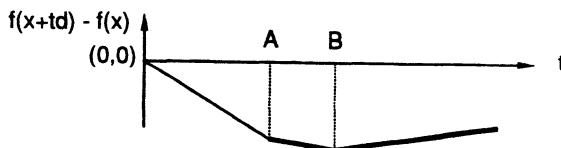
implying that the projection of the origin onto the line-segment  $[\nabla f_1(x), \nabla f_2(x)]$  is just  $\nabla f_1(x)$ . Figure 2.2.2 shows this direction  $d$ . Observe that the straight line  $x + \mathbb{R}d$  passes above the origin; the reason is that  $3/2$ , the slope of  $\nabla f_1(x)$ , is smaller than  $3$ ,



**Fig. 2.2.2.** Non-convergence of steepest descent

the slope of  $L_+$ . The one-dimensional function  $t \mapsto f(x + td)$  is piecewise affine with two kinks  $A$  and  $B$ .

Now take a “reasonable” stepsize along this direction; for example the optimal stepsize. Figure 2.2.3 shows that  $f$  is decreasing along the segment  $AB$ , and that the optimal stepsize leads  $x$  to  $B$ , i.e. on  $L_-$ ; this can be confirmed by simple calculations. By symmetry, the same algorithm starting from this new  $x \in L_-$  will end up with a next iterate on  $L_+$ . Clearly enough, the resulting sequence will oscillate forever between  $L_+$  and  $L_-$ , and converge to the nonoptimal point  $(0, 0)$ . The algorithm is subject to zigzags.



**Fig. 2.2.3.** Objective-values along steepest descent

**Remark 2.2.1** The whole idea of our construction is that, for each iterate, the direction of search passes above the origin; the next iterate, given by the line-search, will then have  $\eta > 0$ . For the direction issued from  $x = (\xi, \eta)$  to pass above the origin, it suffices to have  $\eta$  not too small, namely

$$5\eta > 2|\xi| > 0. \quad (2.2.2)$$

Up to now, we have required that  $f_{\pm 1}$  be active at each iterate, i.e.

$$\eta \geq 3|\xi| (> 0) \quad (2.2.3)$$

which implies (2.2.2). If (2.2.3) did not hold, the direction would become  $(5, -2)$  or  $(-5, -2)$ . Redrawing Fig. 2.2.2 accordingly, we see that the optimal stepsize would lead the next iterate to  $A$ . The next direction would be vertical, pointing directly down to the optimal set; the counter-example would disappear.

It is important to check that the counter-example is not affected by perturbations: otherwise, one could object that it is too much adhoc.

- Suppose first that Step 3 in Algorithm 2.1.5 produces an optimal stepsize at each iteration. Then the first iterate  $x_1$  does not have to lie exactly on  $L_{\pm}$  but simply *above*  $L_{\pm}$ : it suffices to have (2.2.3) at the first iteration. Then we have seen that  $x_2, x_3, \dots$  are all on  $L_{\pm}$ , implying a fortiori (2.2.2) at all iterations  $k \geq 2$ . The situation has not really changed with respect to Fig. 2.2.2.
- Still assuming that (2.2.2) holds at the first iteration, consider now a line-search designed according to the principles of §II.3: for example Wolfe’s line-search II.3.3.1. Look again at Fig. 2.2.3, and visualize on it the two tests (II.3.2.3) and (II.3.2.4); at least if  $m'$  is small, the line-search produces a point slightly beyond  $B$ . Then the reader can convince himself that the counter-example still works, providing that  $t_k$  deviates from optimality by an amount small enough, but strictly positive when the direction happens to be  $\nabla f_{\pm 2}(x_k)$ . In other words, the next iterate must never be on  $A$ , and must not be too far beyond  $B$ .
- Finally, observe that non-convergence still occurs under perturbations of the direction, which does not have to be steepest. For example, as long as (2.2.2) holds, we can take any subgradient in  $\partial f(\xi, \eta)$ : its opposite is downhill and does not suppress the zigzag problem.

□

Our counter-example is worth meditating, because it suggests the profound causes for non-convergence of the steepest-descent method.

- (i) In our present case of the Euclidean norming, the steepest-descent direction at  $x$  is a well-defined function of  $x$ ; call it  $d(x)$ . Clearly enough,  $x \mapsto d(x)$  is not continuous. In our example, when  $x$  varies,  $d(x)$  jumps between the five values (neglecting the normalization)

$$(\pm 5, -2), \quad (\pm 2, -3), \quad (0, -3). \quad (2.2.4)$$

When proving convergence of a steepest-descent method, such as in §II.2.2, it is usually crucial to establish a continuity property: if  $x_k \rightarrow x^*$ , we need  $d_k$  to converge to  $d^*$ , the steepest-descent direction at  $x^*$ . This cannot hold here: observe for example that, in our construction,  $x_k \rightarrow 0$ ; but  $d_k$  certainly does not tend to  $(0, -3)$ , which is  $d(0)$ . This gives a mathematical explanation of the zigzags.

- (ii) Another explanation is “infinite short-sightedness”, which is just the same word as discontinuity, but more intuitive. Among the five possibilities of (2.2.4), only the direction  $(0, -3)$  is able to drive an iterate to the optimal set (unless one starts from an  $x = (\xi, \eta)$  with fairly large  $|\xi|$ ). This  $(0, -3)$  prevails only if  $\xi = 0$ : in order to get it, one must “see” a pair of opposite functions in  $J(x)$ , say  $f_1$  and  $f_{-1}$  (or  $f_2$  and  $f_{-2}$ ). Unfortunately, no matter how close  $\xi$  is to zero, this simultaneous view is impossible:  $f_1$  and  $f_{-1}$ , say, cannot be both active at  $(\xi, \eta)$  unless  $\xi = 0$ .

Take for example an iterate  $x_k = (\xi_k, \eta_k)$  with  $\xi_k > 0$ , and  $J(x_k) = \{1, 2\}$ , say. We know that  $f_{-1}$  will be active at the next iterate (which can be fairly close to  $x_k$ ). Nevertheless, the algorithm does not have our higher view of the situation and does not guess that  $f_{-1}$  should somehow be taken into account *already* when computing the direction at  $x_k$ . The trouble is that the steepest-descent mechanism does not anticipate at  $x_k$  what is going to happen at  $x_{k+1}$ .

- (iii) We have mentioned in §II.2 that in classical optimization of smooth functions, steepest-descent schemes, and any kind of first-order methods, are bad because of their slow convergence. When the objective function becomes more and more ill-conditioned – i.e. when  $\nabla f(x)$  varies more and more rapidly with  $x$  – this slowness becomes worse and worse. At the limit, when  $\nabla f(x)$  varies “infinitely rapidly” – i.e. when it becomes frankly discontinuous, as in the present context – the method becomes frankly non-convergent. This point will be illustrated in §3.3.
- (iv) Chances to obtain convergence for the function of (2.2.1) could be recovered, provided that the stepsize were adequately chosen (while keeping the steepest-descent direction). In Fig. 2.2.4, such an adequate  $t_k$  is the “shortest significant” one, i.e. the smallest  $t > 0$  for which  $J(x + td) \not\subset J(x)$ . Starting from  $x_1$  with  $J(x_1) = \{2\}$ , we go to  $x_2 = A$  where the active set is  $\{1, 2\}$ ; one more similar iteration and the active set becomes  $\{-1, +1\}$  yielding a vertical steepest-descent direction and ending the game. This trick is a key to the development of *pivoting algorithms* for the minimization of piecewise affine functions, see §3.4 below.

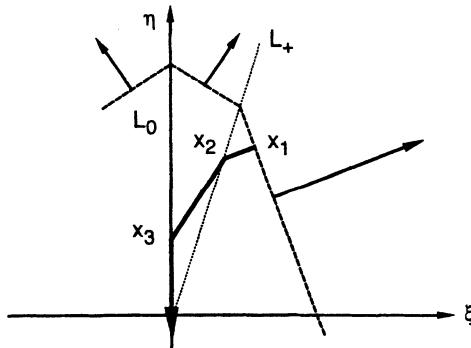


Fig. 2.2.4. With shorter steps, steepest descent should converge

## 2.3 Connection with Nonlinear Programming

Let us come back to our original minimax problem (2.0.1), written as

$$\begin{aligned} \min r \quad & (x, r) \in \mathbb{R}^n \times \mathbb{R}, \\ f_j(x) \leq r \quad & \text{for } j = 1, \dots, p. \end{aligned} \tag{2.3.1}$$

Both problems are “equivalent”, in the sense that  $\bar{x}$  solves (2.0.1) if and only if the pair  $(\bar{x}, f(\bar{x}) =: \bar{r})$  solves (2.3.1). Now, (2.3.1) is an ordinary convex constrained minimization problem with smooth data, which we can write in the more general form

$$\min \{F(z) : c_j(z) \leq 0 \text{ for } j = 1, \dots, p\}. \tag{2.3.2}$$

Here  $z$ , standing for  $(x, r)$ , is a variable in some space  $Z$ , standing for  $\mathbb{R}^{n+1}$ ; the function  $F$  stands for  $r = z^{n+1}$  and

$$c_j(z) = f_j(x) - r = f_j(z^1, \dots, z^n) - z^{n+1} \quad \text{for } j = 1, \dots, p. \quad (2.3.3)$$

Under these conditions, a natural question is: what if we apply to (2.0.1) the methods of ordinary nonlinear programming via (2.3.1) – (2.3.3)? This question has several aspects.

**(a) Minimality Conditions** Take an arbitrary  $x_0 \in \mathbb{R}^n$ ; then the associated point  $z_0 := (x_0, f(x_0) + 1) \in \text{epi } f$  clearly indicates that (2.3.1) – (2.3.2) satisfies the strong Slater assumption (VII.2.3.1). Forming the Lagrange function

$$L(x, r, \mu) := r + \sum_{j=1}^p \mu_j [f_j(x) - r],$$

we can then apply the minimality conditions of §VIII.2... and they turn out to be nothing but  $0 \in \partial f(x)$ , the minimality condition of Proposition 2.1.2. In fact, these conditions are:  $\bar{z} = (\bar{x}, \bar{r})$  solves (2.3.1) if and only if there are multipliers  $\mu_j$  satisfying

$$\mu_j \geq 0 \quad \text{for } j = 1, \dots, p \quad (2.3.4)$$

$$f_j(\bar{x}) \leq \bar{r} \quad \text{for } j = 1, \dots, p \quad (2.3.5)$$

$$(0 \in \mathbb{R}^n, 1) + \sum_{j=1}^p \mu_j (\nabla f_j(\bar{x}), 1) = (0 \in \mathbb{R}^n, 0) \quad (2.3.6)$$

$$\mu_j [f_j(\bar{x}) - \bar{r}] = 0 \quad \text{for } j = 1, \dots, p. \quad (2.3.7)$$

Furthermore, common sense tells us that  $\bar{r}$  is actually  $f(\bar{x})$ : a larger value could certainly not be minimal in (2.3.1). Then (2.3.7) implies that the event  $\mu_j > 0$  can occur only for  $j \in J(\bar{x})$ . From (2.3.4) and the  $r$ -part of (2.3.6), we deduce that the  $\mu_j$ 's actually form a set of convex multipliers. Finally, the  $x$ -part of (2.3.6) means that the corresponding combination of gradients is 0. Altogether, we just obtain Proposition 2.1.2.

**(b) Projected Gradients in Nonlinear Programming** For our constrained minimization problem with smooth data (2.3.2), call

$$C := \{z \in Z : c_j(z) \leq 0 \text{ for } j = 1, \dots, p\}$$

the feasible set and let  $\langle \cdot, \cdot \rangle$  be a scalar product in  $Z$ . A central numerical tool is then the linearization of (2.3.2) around a given  $z \in C$ :

$$\left| \begin{array}{l} \min \langle \nabla F(z), \zeta \rangle \\ \zeta \in T_C(z), \\ v(\zeta) = 1, \end{array} \right. \quad (2.3.8)$$

where  $T_C(z)$  is the tangent cone to  $C$  at  $z$ , and  $v$  is a norm on  $Z$ . The motivation comes directly from §II.2.2 and Definition 2.1.2: the solutions of (2.3.8) are the tangent (normalized) directions that are most downhill with respect to the objective function  $F$ .

In practice, the only interesting case here is as follows: a Slater assumption holds for  $C$ ,  $z$  is not optimal in (2.3.2), and  $v$  is a “quadratic norm” associated with some symmetric positive definite operator  $Q : Z \rightarrow Z$ . In terms of the set  $J(z)$  of active constraints at  $z$ , (2.3.8) is written (we use positive homogeneity,  $\kappa$  is a positive number as in §1)

$$\begin{cases} \min \langle\langle \nabla F(z), \zeta \rangle\rangle \\ \langle\langle \nabla c_j(z), \zeta \rangle\rangle \leq 0 \quad \text{for } j \in J(z), \\ \frac{1}{2} \langle\langle Q\zeta, \zeta \rangle\rangle \leq \kappa. \end{cases} \quad (2.3.9)$$

**Proposition 2.3.1** *With the notations and assumptions introduced above, call  $\hat{s}$  the projection of  $-\nabla F(z)$  onto  $N_C(z)$  for the quadratic norm associated with  $Q^{-1}$ ; in other words,  $\hat{s}$  is the unique solution of*

$$\min \{ \langle\langle \nabla F(z) + s, Q^{-1}[\nabla F(z) + s] \rangle\rangle : s \in N_C(z) \}. \quad (2.3.10)$$

The unique solution  $\hat{s}$  of (2.3.9) is then collinear to  $-Q^{-1}[\nabla F(z) + \hat{s}]$ . If  $Q = I$ , this is the projection of  $-\nabla F(z)$  onto  $T_C(z)$ .

PROOF. The proof goes as in Case 3 of §1.4. Using  $\zeta_0 = 0$  shows that the weak Slater assumption holds in (2.3.9) and we take the Lagrange function

$$L(\zeta, \mu) = \langle\langle \nabla F(z) + \sum_{j \in J(z)} \mu_j \nabla c_j(z), \zeta \rangle\rangle + \frac{1}{2} \mu_0 \langle\langle Q\zeta, \zeta \rangle\rangle.$$

Here  $\mu_0$  is going to be positive and the  $\mu_j$ 's are nonnegative for  $j \in J(z)$ . We set

$$s(\mu) := \sum_{j \in J(z)} \mu_j \nabla c_j(z),$$

which describes  $N_C(z)$  when the  $\mu_j$ 's describe  $\mathbb{R}^+$ .

The Lagrangian  $L$  must be minimal with respect to  $\zeta$  (Proposition VII.3.1.4), which gives

$$\hat{\zeta} = -\frac{1}{\mu_0} Q^{-1}[\nabla F(z) + s(\mu)];$$

the precise value of  $\mu_0$  is absorbed by  $\kappa$ . As for the dual problem, which gives the multipliers – i.e.  $\hat{s}$  –, straightforward calculations show that it is just (2.3.10).

When  $Q$  is the identity operator of  $Z$ , i.e. when the projection is done for the Euclidean norm  $\langle\langle \cdot, \cdot \rangle\rangle^{1/2}$ , we apply §III.3.2 to obtain the stated property:

$$-\nabla F(z) = \hat{s} + \mu_0 \hat{\zeta} \quad \text{and} \quad \langle\langle \hat{s}, \hat{\zeta} \rangle\rangle = 0.$$

□

The above result explains the terminology of *projected gradient* for the optimal solution of (2.3.9). It also gives another motivation for (2.3.8): the direction is taken tangent to  $C$ , but as close as possible to the desired direction  $-\nabla F(z)$ . Observe that the projected gradient is 0 if and only if  $z$  solves (2.3.2): this is (ii') and (iii) in Theorem VII.1.1.1.

**Remark 2.3.2** When the solution  $\hat{\zeta}$  of (2.3.9) is nonzero, it can be used for a line-search yielding the next iterate under the form  $z + t\hat{\zeta}$ . However, the resulting method is delicate for two reasons:

- If there is an active constraint at  $\hat{\zeta}$ , i.e. if  $\langle \nabla c_j(z), \hat{\zeta} \rangle = 0$  for some  $j \in J(z)$ , the direction of search may not be feasible in case the corresponding constraint-function  $c_j$  is non-affine:  $z + t\hat{\zeta} \notin C$  whenever  $t > 0$ .

Then there will be some difficulty to find a next iterate in  $C$ . This explains that the use of the projected gradient for minimization algorithms is generally limited to affine constraints only.

- If all the constraints  $c_j$  are affine, the line-search can be devised according to the principles of §II.3; the only modification being to force the stepsize  $t$  to satisfy  $z + t\hat{\zeta} \in C$ . The resulting method is however *non-convergent*. It suffers the short-sightedness evoked in (ii), at the end of §2.2: the next active set  $J(z + t\hat{\zeta})$  is ignored by the direction computed at the present iterate  $z$ .  $\square$

**(c) Projected Gradients and Steepest-Descent Directions** Now let us transcribe the projected-gradient problem (2.3.8) in our present minimax context of (2.3.1). Towards this end, we need to specify the scalar product and the norming in  $Z$ . Because  $Z$  is the product space  $\mathbb{R}^n \times \mathbb{R}$ , it is natural to take

$$\langle z, \zeta \rangle = \langle (x, r), (d, \rho) \rangle := \langle x, d \rangle + \lambda_0 r \rho, \quad (2.3.11)$$

where  $\langle \cdot, \cdot \rangle$  is our scalar product in  $\mathbb{R}^n$ , and  $\lambda_0 > 0$  is arbitrary. Observing that the objective function in (2.3.2) has the gradient  $\nabla F(z) = (0, 1) \in \mathbb{R}^{n+1}$ , the linearized objective function in (2.3.8) is then  $\lambda_0 \rho$ , or more simply  $\rho$ :  $\lambda_0$  is irrelevant. Likewise, we take as norming

$$v(d, \rho) := \|d\| + \lambda|\rho|, \quad (2.3.12)$$

where  $\|\cdot\|$  is a norm on  $\mathbb{R}^n$  and  $\lambda > 0$  is arbitrary. In summary, the “steepest-descent tangent directions” associated with (2.3.1) are the solutions of

$$\begin{cases} \min \rho \\ \langle \nabla f_j(x), d \rangle \leq \rho \quad \text{for } j \in J(x), \\ \|d\| + \lambda|\rho| = 1; \end{cases}$$

or, in a more formal writing (allowing more general than max-functions)

$$\begin{cases} \min \rho \\ f'(x, d) \leq \rho, \\ \|d\| + \lambda|\rho| = 1. \end{cases} \quad (2.3.13)$$

This last problem is right in the framework of §1.

**Proposition 2.3.3** *No matter how  $\lambda > 0$  is chosen in the norming (2.3.12), the solution-set of (1.1.3) and the  $d$ -part of the solution-set of (2.3.13) are collinear.*

PROOF. Call  $\rho_\lambda$  the optimal objective-value in (2.3.13) and set

$$\kappa := 1 - \lambda|\rho_\lambda| \geq 0.$$

We claim  $\kappa > 0$ : indeed,  $\rho_\lambda = 1/\lambda$  would imply  $d = 0$  for any feasible  $(d, \rho_\lambda)$ , and a better objective-value could be obtained with a small enough perturbation of  $d = 0$ . A posteriori, we can replace (2.3.13) by

$$\left| \begin{array}{l} \min \rho \\ f'(x, d) \leq \rho, \\ \|d\| = \kappa, \end{array} \right. \quad (2.3.14)$$

whose solution-set clearly coincides with the  $d$ -part of the solution-set of (2.3.13). Then apply Proposition 1.1.5: if (2.3.14) has a wrong  $\kappa$ , its solution-set is just multiplied by a positive factor.  $\square$

Let us conclude: the steepest-descent directions for (2.0.1) are the “steepest-descent tangent directions” for (2.3.1); and this holds for any scalar product and norming in  $\mathbb{R}^n$ , and any corresponding scalar product and norming satisfying (2.3.11), (2.3.12). As a result, the steepest-descent Algorithm 2.1.5 is just a special form of projected-gradient algorithm, applicable when (2.3.2) has the special form (2.3.1).

To interpret the line-search, take an optimal solution  $(\hat{d}, \hat{\rho}) = \hat{\zeta}$  of (2.3.13) at the point  $(x, r) = z$ . In a “pure” projected-gradient method, one would take the next iterate along the half-line

$$x(t) = x + t\hat{d}, \quad r(t) = r + t\hat{\rho}, \quad t \geq 0.$$

Rather, Algorithm 2.1.5 takes the next iterate along the “vertical” curve

$$x(t) = x + t\hat{d}, \quad r(t) = f(x(t)).$$

**Remark 2.3.4** Some mechanism has to be invented to force convergence of steepest-descent schemes – or equivalently projected gradients. This will be the object of subsequent chapters (starting from Chap. XIII). The basis for such mechanisms is, one way or another, a redefinition of the tangent cone, for example via a modification of the index-set  $J(x)$  from (2.1.1).

As already mentioned,  $J(x)$  is for most  $x$  a singleton. For a reader familiar with numerical calculations, the definition of  $J(x)$ , based on equality in (2.1.1), does not make real sense. For one thing, finite arithmetic is usually used: one should at least take those  $j$  satisfying something like

$$f_j(x) \geq f(x) - \eta,$$

where  $\eta > 0$  accounts for the computing accuracy. This aspect should be kept in mind when speaking of active sets and cones of tangent directions.  $\square$

Another interesting consequence of Proposition 2.3.3 is that the two fields of nonsmooth optimization and nonlinear programming are closely related, via the equivalence between (2.0.1) and (2.3.1). Each of these two fields can thus benefit from improvements in the other.

For example, it has already been mentioned on several occasions (§II.2, and also at the end of §1.1) that steepest-descent schemes are most inefficient. Now, there are known improvements of (2.3.8) to solve (2.3.1); they involve: (i) redefinitions of the concept of tangency (cf. Remark 2.3.4), and (ii) a clever choice of the norm  $\|\cdot\|$  (remember the beginning of §II.2.2). These improvements can serve as basis for solving nonsmooth optimization problems more efficiently. We will see that (i) can be readily transcribed to the context of nonsmooth optimization; as for (ii), it is unfortunately much more delicate.

### 3 The Practical Value of Descent Schemes

Section 2 was mainly devoted to the zigzagging phenomenon, common to all steepest-descent methods. Another problem was mentioned at the end of §1.1, namely that the practical implementation of such methods could be difficult. The full subdifferential had to be computed, and one had to hope that it was a closed convex polyhedron, or an ellipsoid; see §1.4. The aim of the present section is to show that, in many situations, such a computation is not convenient.

#### 3.1 Large Minimax Problems

Take again the max-function (2.0.1) but suppose  $p$  is a large integer, say in the  $10^6$ -range. Then the mere task of computing the active set  $J(x)$  is unreasonable, not even mentioning the projection problem (2.1.6). We illustrate instances of such large minimax problems with the exact penalty technique: consider an ordinary nonlinear programming problem

$$\left| \begin{array}{l} \min F(x) \\ c_j(x) \leq 0 \quad \text{for } j = 1, \dots, p, \end{array} \right. \quad (3.1.1)$$

with smooth objective- and constraint-functions  $F$  and  $c_j$ , but an extremely large number  $p$  of constraints. Known methods for constrained optimization become impractical in this situation; accordingly, the penalty idea is to transform the problem by aggregating the constraints into the objective function; see Chap. VII if necessary.

The approach with an  $\ell_\infty$ -penalty is standard: the penalty coefficient  $\pi$  is chosen (“large enough”) and the following function is minimized without constraints:

$$\mathbb{R}^n \ni x \mapsto F(x) + \pi \max\{0, c_1(x), c_2(x), \dots, c_p(x)\}. \quad (3.1.2)$$

Thus, one has a genuine minimax problem to solve, with a max operation involving  $p+1$  terms.

**Remark 3.1.1** Among the many other exact penalties, we rather considered in §VII.3.2 the  $\ell_1$ -approach in which, instead of (3.1.2), it is the function

$$\mathbb{R}^n \ni x \mapsto F_\pi(x) := F(x) + \pi \sum_{j=1}^p \max\{0, c_j(x)\} \quad (3.1.3)$$

that was minimized. This function can be put in the minimax framework of §2:

$$F_\pi(x) = F(x) + \pi \max\{\varepsilon_1 c_1(x) + \dots + \varepsilon_p c_p(x) : \varepsilon_j \in \{0, 1\}\},$$

so  $F_\pi$  is a max of  $2^p$  functions. To characterize  $\partial F_\pi$  from this last expression, denote by

$$J(x) := \{j : c_j(x) = 0\}$$

the set of active indices at a given  $x$ , and by

$$s_0(x) := \nabla F(x) + \pi \sum_{\{j: c_j(x) > 0\}} \nabla c_j(x)$$

the “smooth part” of the differentiation. The subdifferential of  $F_\pi$  at  $x$  is therefore the convex hull of  $2^{|J(x)|}$  points:

$$\partial F_\pi(x) = s_0(x) + \pi \operatorname{co} \left\{ \sum_{j \in J(x)} \varepsilon_j \nabla c_j(x) : \varepsilon_j \in \{0, 1\} \right\}.$$

Because  $\partial F_\pi(x)$  has exponentially many extreme points, it is more conveniently characterized directly from (3.1.3): applying the calculus rules of §VI.4 gives

$$\partial F_\pi(x) = \{s_0(x)\} + \pi \sum_{j \in J(x)} [0, 1] \nabla c_j(x),$$

which places oneself in Case 2 of §1.4. Computing a steepest-descent direction is now a convex minimization problem with  $|J(x)|$  variables  $\alpha_j$ :

$$\begin{cases} \min \left\| s_0(x) + \pi \sum_{j \in J(x)} \alpha_j \nabla c_j(x) \right\|^* \\ 0 \leq \alpha_j \leq 1 \quad \text{for } j \in J(x). \end{cases}$$

Another observation concerns the constrained minimization problem playing the role of (2.3.1): to minimize  $F_\pi$  is equivalent to

$$\begin{cases} \min \left[ F(x) + \pi \sum_{j=1}^p r_j \right] & x \in \mathbb{R}^n, \quad r \in \mathbb{R}^p, \\ r_j \geq 0, \quad r_j \geq c_j(x) & \text{for } j = 1, \dots, p. \end{cases} \quad (3.1.4)$$

We leave it as an exercise to reproduce on (3.1.4) the development of §2.3: work out the projected-gradient problem, and interpret it in terms of §1, via a result playing the role of Proposition 2.3.3.  $\square$

Whether the chosen variant is (3.1.2) or (3.1.3), one must admit that  $J(x)$  is potentially an untractably large set when  $p$  is really large in (3.1.1) (the argument that  $J(x)$  is for most  $x$  a singleton should not be taken too seriously, remember Remark 2.3.4). We have here a first example suggesting that effective computation of the whole subdifferential may not be a reasonable task.

The dimensionality argument becomes really critical when the max operation in (2.0.1) has itself a combinatorial aspect. Consider the problem of minimizing the convex function whose value at  $x$  is

$$f(x) := \max \left\{ \sum_{i=1}^q y_i f_i(x) : Ay = b, \quad y_i \geq 0 \text{ for } i = 1, \dots, q \right\}, \quad (3.1.5)$$

where each  $f_i$  is smooth and convex,  $A$  is some  $m \times q$  matrix, and  $b \in \mathbb{R}^m$ .

Despite the appearances, minimizing this  $f$  is a finite minimax problem: the underlying max-operation consists of maximizing a linear function (in  $y$ ) over a closed convex polyhedron. Assuming  $f$  finite everywhere and remembering the end of Chap. V, this maximization can be restricted to the (finitely many) *extreme points* of the same polyhedron. In a word,  $f$  has just the form (2.0.1), with  $j$  indexing these extreme

points. But who can characterize all of them, or even count them? The only thing that can be done in practice is to compute just one maximal  $y$ , but certainly *not all*.

In this situation, a numerical process (a computer program, implementing a suitable linear programming algorithm) is usually at hand which, given  $x$ , computes a maximal  $y$ , say  $\bar{y}$ ; remember the black box (U1) of Fig. II.1.2.1. Then the vector  $s := \sum_{i=1}^q \bar{y}_i \nabla f_i(x)$  is in  $\partial f(x)$ . There is no hope to compute the others, and there is no hope to compute a steepest-descent direction. We do not even mention the larger set alluded to in Remark 2.3.4.

### 3.2 Infinite Minimax Problems

A function like (3.1.5) is on the fringe between the finite minimax problem (2.0.1) and the case with infinitely many indices:

$$f(x) := \max \{h(x, y) : y \in Y\} \quad (3.2.1)$$

where  $h$  is convex in  $x$ , and smooth on the compact set  $Y$ .

Among such problems, are the frequently encountered *semi-infinite programs* (see §VI.5.3):

$$\left| \begin{array}{l} \min F(x) \\ g(x, t) \leq 0 \quad \text{for all } t \in T, \end{array} \right. \quad (3.2.2)$$

where  $T$  is a compact interval of  $\mathbb{R}$ . The latter does not fit exactly in (3.2.1) but the reader should observe that minimizing a function – as in (3.2.1) – and imposing an inequality constraint – as in (3.2.2) – are twin brothers. To satisfy  $g(\cdot, \cdot) \leq 0$  is usually achieved by means of decreasing  $g$ ! In fact, the constraint-function of interest in (3.2.2) is

$$x \mapsto c(x) := \max_{t \in T} g(x, t).$$

Observe also that (3.2.1) and (3.2.2) become fully equivalent if the semi-infinite program is attacked via a penalty technique such as in §3.1: the function

$$F(x) + \pi \max_{t \in T} g(x, t)$$

has undoubtedly the form (3.2.1) (note that the  $\ell_1$  penalty – or rather  $L_1$  – is not so convenient here, as it would involve an integration in the space  $T$ ).

Clearly enough, it is even “more impossible” in (3.2.1) than in (3.1.5) to compute the whole subdifferential  $\partial f(x)$ , which amounts to computing the whole, potentially infinite, set of maximal  $y$ ’s at  $x$ . On the other hand, computing *some* subgradient – of  $f$  in (3.2.1), of  $c$  in (3.2.2) – can also be hard enough: the underlying maximization has no reason to be easy. We simply observe, however, that it has to be performed anyway, just to compute the objective- or the constraint-values. Once this is done, a subgradient is available “for free”: differentiate  $h(\cdot, y)$  or  $g(\cdot, t)$  for the  $y$  or  $t$  just obtained (Lemma VI.4.4.1).

At any rate, problems of the type (3.2.1) or (3.2.2), with  $f$  or  $c$  relatively easy to compute, cover an extremely large range of practical applications. It will be seen in Chap. XII that such is the case in the (very important) field of *decomposition*.

Other instances of infinite minimax problems are the (also important) cases of maximal eigenvalues. It has been seen in §VI.5.1 that the maximal eigenvalue  $\lambda_1(M)$  of a varying symmetric matrix  $M$  is convex and that its subdifferential, in an appropriate Euclidean space, is

$$\partial\lambda_1(M) = \text{co} \{ uu^\top : u \text{ normalized eigenvector associated with } \lambda_1 \}.$$

To compute the full subdifferential, one must first know the multiplicity of the largest eigenvalue. In practice, this implies the computation of the full spectrum of  $M$ . Let  $m > 1$  be the multiplicity of  $\lambda_1$  (the case  $m = 1$  is easy) and suppose that an orthonormal system of corresponding eigenvectors  $v_1, v_2, \dots, v_m$  is known; such is usually the case after the spectral decomposition of  $M$  has been performed. Then, setting for  $\alpha \in \mathbb{R}^m$

$$w(\alpha) := \sum_{k=1}^m \alpha_k v_k,$$

we have

$$\partial\lambda_1(M) = \text{co} \{ w(\alpha)w^\top(\alpha) : \alpha^\top \alpha = 1 \}.$$

We refer to §VI.5.1 for a characterization of this set, and the computation of the directional derivative  $\lambda'_1(M, P)$ . Computing *one* subgradient at a given  $M$  amounts to computing the maximal eigenvalue of  $M$  – without caring whether it is multiple – and one associated eigenvector. This is usually a simpler task.

### 3.3 Smooth but Stiff Functions

The previous sections dealt with problems in which computing the subdifferential was difficult, or even impossible. In another field of applications, computing it is simply meaningless. This concerns objective functions whose gradient varies rapidly, although continuously.

In theory, there are excellent algorithms to minimize such functions; in particular of Newton-type, see §II.2.3. Unfortunately, there is no clear-cut between functions that are smooth (whence in the field of application of such algorithms) and functions that are not (whence requiring methods from nonsmooth optimization). Between these two classes, there is a rather fuzzy boundary of *stiff functions*, for which it is not clear what class of methods is better suited.

A nonsmooth function  $f$  can be regarded as the limiting case of a  $C^2$  function, say  $g$ , whose second derivatives grow unboundedly at some points (the kinks of  $f$ ). In intuitive words, the Hessian matrix of  $g$  at such a point  $x$  has some very large eigenvalues: those corresponding to eigenvectors parallel to  $\text{aff } \partial f(x)$ . On the other hand, the gradient of  $g$  stays bounded: the limiting function  $f$  satisfies a Lipschitz property. For  $\varepsilon > 0$ , think of the example

$$\mathbb{R}^2 \ni x = (\xi, \eta) \mapsto g(x) := \sqrt{\xi^2 + \varepsilon} + \eta \simeq |\xi| + \eta =: f(x).$$

We have  $\frac{\partial^2 g}{\partial \xi^2}(0) = \varepsilon^{-1/2}$  and  $\nabla^2 g(0)$  has the eigenvalue  $\varepsilon^{-1/2}$ , with the associated eigenspace  $\mathbb{R} \times \{0\}$ , parallel to  $[-1, +1] \times \{1\} = \partial f(0)$ .

**Remark 3.3.1** A popular example of stiff functions comes from the quadratic penalty. To solve (3.1.1), consider instead of (3.1.3) the following variant:

$$F(x) + \pi \sum_{j=1}^p [c_j^+(x)]^2,$$

which is definitely smooth (although not  $C^2$ ). It is reasonable to believe that its minima approach solutions of (3.1.1) when  $\pi \rightarrow \infty$ . Functions of this type do not illustrate the kind of stiffness that we have in mind, though: when  $\pi \rightarrow \infty$ , their gradients are unbounded at all nonfeasible points. By contrast, observe that our function  $g$  above has the gradient  $\nabla g(\xi, \eta) = \left( \frac{\xi}{(\xi^2 + \varepsilon)^{1/2}}, 1 \right)$ , which stays in  $[-1, +1] \times \{1\}$ .  $\square$

Then the question is: when is a (finite) eigenvalue so large that it looks infinite to an optimization algorithm? In other words: when is a smooth function so stiff that an algorithm tailored for smooth functions will become inefficient? Answering this question is quite complex. The following experiment illustrates how vague the class of stiff functions is; at the same time, it presents a possible technique of *regularization* for finite minimax problems. The objective function in (2.0.1) can be written

$$f(x) = \max_{\alpha \in \Delta_p} \sum_{j=1}^p \alpha_j f_j(x), \quad (3.3.3)$$

where  $\Delta_p$  is the unit simplex. Now, take  $\pi > 0$  and set

$$\mathbb{R}^n \times (\mathbb{R}_*^+)^p \ni (x, \alpha) \mapsto \varphi^\pi(x, \alpha) := \sum_{j=1}^p \alpha_j f_j(x) + \pi \sum_{j=1}^p \log \alpha_j.$$

For small  $\pi > 0$ ,  $\varphi^\pi$  approximates the maximand in (3.3.3), and the function

$$x \mapsto f^\pi(x) := \max \{ \varphi^\pi(x, \alpha) : \sum_{j=1}^p \alpha_j = 1 \} \quad (3.3.4)$$

approximates  $f$ : it can indeed be proved that, for all  $x$ ,

$$f^\pi(x) \uparrow f(x) \quad \text{when } \pi \downarrow 0. \quad (3.3.5)$$

**Remark 3.3.2** The computation of  $f^\pi(x)$  for fixed  $x$  is not difficult; observe in particular that  $\varphi^\pi(x, \cdot)$  is strictly concave. Call  $\lambda^\pi(x)$  the Lagrange multiplier associated with the constraint  $\sum_{j=1}^p \alpha_j = 1$  in (3.3.4); the maximality conditions for  $\alpha$  readily give

$$\frac{1}{\alpha_j} = -\frac{f_j(x) + \lambda^\pi(x)}{\pi}, \quad \sum_{j=1}^p \alpha_j = 1.$$

The multiplier  $\lambda^\pi(x)$  can therefore be computed via a Newton method to solve the equation in  $\lambda$

$$\sum_{j=1}^p \frac{\pi}{f_j(x) + \lambda} + 1 = 0. \quad (3.3.6) \quad \square$$

In contrast with (3.3.3), the  $\alpha$ -problem (3.3.4) making up  $f^\pi$  has a unique maximal solution, so the resulting  $f^\pi$  is now differentiable. This can be seen directly: applying the implicit function theorem to (3.3.6), we see that  $\lambda^\pi(\cdot)$  is indeed  $C^\infty$  if the  $f_j$ 's are such. In the present context, we now have to minimize with respect to  $x$  a smooth function  $f^\pi(x)$ , instead of the nonsmooth  $f(x)$ .

For an illustration, we take the following minimax problem:

**Test-problem 3.3.3 (MAXQUAD)** Equip  $\mathbb{R}^n$  with the usual dot-product and take in (2.0.1)

$$f_j(x) := \frac{1}{2}x^\top A_j x + b_j^\top x + c \quad \text{for } j = 1, \dots, p$$

where each  $A_j$  is a symmetric positive definite  $n \times n$  matrix and  $b_j$  an  $n$ -vector;  $c$  is a real number. Then  $\nabla f_j(x) = A_j x + b_j$  and each  $f_j$  is strongly convex.

In a specific example called MAXQUAD,  $n = 10$ ,  $p = 5$ ,  $c = 0.8414$  and the minimal value is  $\bar{f} = 0$ . At the (unique) minimum  $\bar{x}$ , the five underlying functions have the characteristics listed in Table 3.3.1:

**Table 3.3.1. Bad scaling in the example MAXQUAD**

j	1	2	3	4	5
$f_j(\bar{x})$	0.	0.	0.	0.	-298.
$\ \nabla f_j(\bar{x})\ $	6.	14.	40.	500.	$10^4$

It will be confirmed below that the fifth function is really special: it is supposedly inactive at  $\bar{x}$  but becomes active at  $\bar{x} + t\nabla f_5(\bar{x})$  very quickly when  $t$  increases; then  $f = f_5$  becomes very steep. On the other hand,  $f$  behaves much more gently when  $t$  decreases.  $\square$

Now, the objective function of MAXQUAD can be approximated by the smooth  $f^\pi$  of (3.3.4), which in turn can be minimized by any “ordinary” method. Table 3.3.2 displays the behaviour of various algorithms, for different values of  $\pi$ : Euclidean steepest descent (§II.2.2), one among the best known implementations of conjugate gradient (§II.2.4) and of quasi-Newton (§II.2.3), and a typical method for convex (nonsmooth) optimization (Algorithm XIV.3.4.2). Each entry of Table 3.3.2 contains the number of iterations to reach three exact digits and, between parentheses, the corresponding number of  $f^\pi$ - and  $\nabla f^\pi$ -evaluations. The last two rows indicate the value of  $f^\pi$  at its minimum  $x^\pi$  and the corresponding value of  $f$ . Note:  $f^\pi(x^\pi) \leq 0 = \bar{f} \leq f(x^\pi)$  because of (3.3.5). All methods were started from the same initial iterate and used the same line-search. The computer had about six exact digits.

This table makes clear enough the danger of believing that a “smooth method” is automatically appropriate for a smooth function.

One may wonder if the range of values of  $\pi$  that are displayed in Table 3.3.2 is significant: for example, could it be that the value  $\pi = 100$  say, is already “very small”, taking into account the regularity properties of  $f^\pi$ ? Actually, some indication is obtained by comparing the optimal values: knowing that the three numbers  $f(x^\pi)$ ,  $f^\pi(x^\pi)$  and  $f(\bar{x}) = 0$  should be close together, we see that  $\pi = 10^{-3}$  is not unreasonably small. In fact,  $\pi$  is homogeneous to  $f$ -values.

**Table 3.3.2.** What is a smooth function?

$\pi$	100	10	1	$10^{-1}$	$10^{-2}$	$10^{-3}$	0.
st. desc.	2(6)	21(35)	30(57)	59(97)	358(487)	$\infty(\infty!)$	failed
conj. grad.	5(15)	10(23)	13(33)	20(50)	77(222)	69(194)	failed
q. Newton	3(8)	15(25)	27(33)	50(73)	91(130)	104(186)	failed
Nonsmooth	3(16)	6(10)	12(20)	15(22)	24(46)	24(54)	17(44)
$f^\pi(x^\pi)$	-1117.	-121.	-14.	-1.68	-.207	-.024	0.
$f(x^\pi)$	184.	8.7	1.6	.17	.014	.001	0.

**Remark 3.3.4** The collapse of the conjugate gradient method and, even more notably, of quasi-Newton methods, can be explained. Both methods construct a quadratic approximation based on observations from the objective function. In view of the nasty behaviour of  $f_5$  alluded to in Table 3.3.1, the present  $f$  is highly unsymmetric. Therefore, the quadratic model cannot approximate well  $f^\pi$  nor  $f$ , apparently even for fairly large values of  $\pi$ . This is confirmed by Table 3.3.3, which reports the same experiments with  $p = 4$  (thus eliminating  $f_5$  from Example 3.3.3), and requiring 4 exact digits instead of 3. The behaviour of the various classical methods is now much more in accordance with their usual behaviour.

Observe in passing the relatively good performance of the “nonsmooth method”, even in smooth cases.  $\square$

**Table 3.3.3.** What is a smooth function? (cont'd)

$\pi$	10	1	$10^{-1}$	$10^{-2}$	$10^{-3}$	$10^{-4}$	0.
st.desc.	6(14)	15(22)	66(95)	894(1014)	$\infty$	$\infty$	failed
conj.grad.	4(9)	7(15)	21(44)	44(107)	94(242)	147(430)	failed
q.Newton	4(6)	7(10)	13(24)	16(27)	23(50)	26(73)	41(163)
Nonsmooth	5(10)	16(31)	19(35)	19(42)	25(57)	22(58)	25(52)
$f^\pi(x^\pi)$	-62.88	-7.183	-.8337	-.0947	-.0150	-.0011	0.
$f(x^\pi)$	6.5	1.3	.16	.015	.0016	.0002	0.

In summary, even for smooth functions, there is room for methods based on convex analysis, which do not assume smoothness. Needless to say, the concept of subdifferential is then irrelevant, and a steepest-descent direction brings nothing with respect to §II.2: the whole theory in §1 becomes useless in this framework.

### 3.4 The Steepest-Descent Trajectory

There is something offensive in the counter-example of §2.2. Imagine that the graph of  $f$  is (a portion of) the surface of the earth and imagine a drop of water rolling very slowly on that surface. If  $x(\tau) \in \mathbb{R}^2$  denotes the position of the drop at time  $\tau$ , the steepest-descent direction  $-\hat{s}(\tau)$  is “tangent” to the curve  $x(\cdot)$ . Now, it is hard to imagine that  $x(\tau)$  converges to anything else than a minimal point, especially if the surface is as simple as polyhedral.

**(a) Continuous Time** Consider first the *differential inclusion*

$$\dot{x}(\tau) \in -\partial f(x(\tau)), \quad x(0) = x_1 \text{ given.} \quad (3.4.1)$$

We will call “solution” to (3.4.1) an absolutely continuous function  $\tau \mapsto x(\tau)$  (see §A.6.1) such that, for almost all  $\tau \geq 0$ , the derivative of  $x$  at  $\tau$  is opposite to some subgradient of  $f$  at  $x(\tau)$ . Admitting that a solution exists in this sense, how about its behaviour for  $\tau \rightarrow +\infty$ ? Does it converge to a minimum point of  $f$  (if any)? We start with a list of fundamental properties.

**Theorem 3.4.1** *With  $f: \mathbb{R}^n \rightarrow \mathbb{R}$  convex, the differential inclusion (3.4.1) has a unique solution  $x(\cdot): [0, +\infty[ \rightarrow \mathbb{R}^n$ . Furthermore:*

- (i) *The function  $\tau \mapsto x(\tau)$  is Lipschitzian from  $[0, +\infty[$  to  $\mathbb{R}^n$ ; it admits a right-derivative  $D_+x(\tau)$  at all  $\tau \geq 0$ , given by*

$$D_+x(\tau) = -\hat{s}(\tau) \quad \text{for all } \tau \geq 0, \quad (3.4.2)$$

*where  $\hat{s}(\tau) = \hat{s}(x(\tau))$  is the orthogonal projection of the origin onto  $\partial f(x(\tau))$ .*

- (ii) *The function  $\tau \mapsto f(x(\tau))$  is convex decreasing from  $[0, +\infty[$  to  $\mathbb{R}$ ; its right-derivative at  $\tau \geq 0$  is  $- \|D_+x(\tau)\|^2 = -\|\hat{s}(\tau)\|^2$ .*

- (iii) *For all nonnegative  $\tau_1$  and  $\tau_2$ , and all  $y \in \mathbb{R}^n$ ,*

$$\frac{1}{2}\|x(\tau_2) - y\|^2 \leq \frac{1}{2}\|x(\tau_1) - y\|^2 - \int_{\tau_1}^{\tau_2} [f(x(\tau)) - f(y)]d\tau. \quad (3.4.3)$$

- (iv) *For all  $T > 0$ ,*

$$f(x(T)) - f(x_1) = - \int_0^T \|\hat{s}(\tau)\|^2 d\tau. \quad \square$$

The proof goes beyond the scope of this book. In a classical Cauchy problem, the continuity of the right-hand side with respect to  $x$  ensures existence of a solution, on some interval  $]0, T[$ . Here, it is essentially the outer semi-continuity of  $\partial f$  (see §VI.6.2) that does the job for existence; and the monotonicity of  $\partial f$  helps making  $T = +\infty$ . Uniqueness, as well as (i) – (iv), come rather easily, via multiplication of both sides of (3.4.1) by appropriate vectors.

We stress an important point in (3.4.2): the equality holds at *all* (rather than almost all)  $\tau \geq 0$ . From a mechanical viewpoint, the dynamics is governed by  $-\hat{s}(\tau)$  at each time  $\tau$ ; the system is “lazy”: it tends to minimize its speed. Thus, (3.4.1) is equivalent to the ordinary differential equation

$$\dot{x}(\tau) = -\hat{s}(\tau), \quad x(0) = x_1 \text{ given;} \quad (3.4.4)$$

the derivative  $\dot{x}(\tau)$  of our absolutely continuous function  $x(\cdot)$  is, for almost all  $\tau \geq 0$ , opposite to the (unique) element of  $\partial f(x(\tau))$  having minimal Euclidean norm. All these properties depend crucially on monotonicity of  $\partial f$ ; they would not be enjoyed by the solutions of  $\dot{x}(\tau) \in \partial f(x(\tau))$ , say. Indeed, differential inclusions hardly tolerate backward integration.

In a word, our differential inclusion (3.4.1) does model the movement of our drop of water: its solution is just the *steepest-descent trajectory* described by (3.4.4). Note also that (ii) implies:

$$\tau \mapsto \|\hat{s}(\tau)\| \text{ is decreasing.} \quad (3.4.5)$$

An important consequence of Theorem 3.4.1 is (as always,  $\bar{f}$  denotes the infimum of  $f$  over  $\mathbb{R}^n$ ):

**Corollary 3.4.2** *Let  $x(\cdot)$  solve (3.4.1) = (3.4.4). When  $\tau \rightarrow +\infty$ ,*

- (i) *the trajectory is minimizing:  $f(x(\tau)) \rightarrow \bar{f}$ ;*
- (ii) *some subgradient tends to 0 if  $\bar{f} > -\infty$ :  $\hat{s}(\tau) \rightarrow 0$ ;*
- (iii) *the trajectory converges to a minimum of  $f$  if there is at least one:  $x(\tau) \rightarrow \bar{x}$ .*

PROOF. Suppose there are  $y \in \mathbb{R}^n$  and  $\varepsilon > 0$  such that

$$f(x(\tau)) - f(y) \geq \varepsilon \quad \text{for all } \tau \geq 0.$$

Then (3.4.3) used with  $\tau_1 = 0$  implies the contradiction:  $\|x(\tau_2) - y\| \rightarrow -\infty$  when  $\tau_2 \rightarrow +\infty$ . As for (ii), it follows directly from Theorem 3.4.1(iv) and (3.4.5).

Now let  $\bar{x}$  minimize  $f$ . To prove (iii), take  $y = \bar{x}$  in (3.4.3):  $f(\bar{x}) - f(x(\tau)) \leq 0$  for all  $\tau \geq 0$  and  $\|x(\tau) - \bar{x}\|^2$  is a decreasing function of  $\tau$ . Thus,  $\{x(\cdot)\}$  is bounded; a cluster point exists, which is optimal in view of (i) and can be called  $\bar{x}$ . Given  $\varepsilon > 0$ , we can find  $\tau_1$  such that

$$\|x(\tau_1) - \bar{x}\|^2 \leq \varepsilon$$

so, using again (3.4.3),

$$\|x(\tau_2) - \bar{x}\|^2 \leq \varepsilon \quad \text{for all } \tau_2 \geq \tau_1. \quad \square$$

Note the special property expressed by (ii): when a curve  $\{x(\cdot)\}$  tends to a minimum of a convex function  $f$ , the subdifferential  $\partial f(x(\cdot))$  may stay away from 0 ( $\partial f$  is not inner semi-continuous!). Here, some subgradient at  $x(\tau)$  tends to 0, which implies first that the convergence of the objective function is fast:

$$f(x(\tau)) - \bar{f} \leq \|\hat{s}(\tau)\| \|\bar{x} - x(\tau)\| = o(\|\bar{x} - x(\tau)\|).$$

It also implies that, for large  $\tau$ ,  $x(\tau)$  stays in a definite region of the space; and this is a result of general interest:

**Proposition 3.4.3** *Let a curve  $\tau \mapsto x(\tau)$  have a limit  $\bar{x}$  for  $\tau \rightarrow +\infty$  and suppose that there exists  $s(\tau) \in \partial f(x(\tau))$  tending to 0. For each  $\tau \geq 0$ , set*

$$d(\tau) := \begin{cases} \frac{x(\tau) - \bar{x}}{\|x(\tau) - \bar{x}\|} & \text{if } x(\tau) \neq \bar{x}, \\ 0 & \text{if } x(\tau) = \bar{x}. \end{cases} \quad (3.4.6)$$

When  $\tau \rightarrow +\infty$ , all nonzero cluster-points of  $\{d(\tau)\}$  are critical directions in the sense of Remark VI.2.2.5:

$$\lim_{\tau \rightarrow +\infty} \text{ext}\{d(\tau)\} \subset N_{\partial f(\bar{x})}(0).$$

PROOF. Let  $d$  be the limit of  $\{d(\tau_k)\}$  for some sequence  $\{\tau_k\}$  tending to  $+\infty$  with  $k$ . If  $d = 0$ , we have nothing to prove; otherwise,  $d_k \neq 0$  for  $k$  large enough.

With  $s_k \in \partial f(x_k)$  tending to 0, take  $s \in \partial f(\bar{x})$  and apply monotonicity of  $\partial f$  in (3.4.6):

$$\langle d(\tau_k), s_k - s \rangle = \frac{\langle s_k - s, x(\tau_k) - \bar{x} \rangle}{\|x(\tau_k) - \bar{x}\|} \geq 0.$$

Letting  $k \rightarrow +\infty$  shows that  $-\langle d, s \rangle \geq 0$ ; hence  $d \in N_{\partial f(\bar{x})}(0)$  since  $s$  was arbitrary in  $\partial f(x)$ .  $\square$

**(b) Piecewise Affine Trajectories** As long as (3.4.4) cannot be solved explicitly, the above development remains dry theory. For a numerical approximation, we have the classical Euler method, in which  $\{x_{k+1}\}$  approximates  $\{x(k\Delta\tau)\}$  for  $k = 1, 2, \dots$ :

$$x_{k+1} = x_k - \Delta\tau s_k \quad \text{with } s_k \in \partial f(x_k).$$

To really approximate the trajectory,  $\Delta\tau$  must be “small”; and to really approximate a limit,  $k\Delta\tau$  must be “large”. Convergence to a minimum will be established if we can mimic Theorem 3.4.1 in a discretized setting, so as to reproduce Corollary 3.4.2. This is the basis for methods of *subgradient optimization*, to be seen later in §XII.4.1.

Here, we content ourselves with the simple situation of a piecewise affine objective function:

$$f(x) = \max \{ \langle s_j, x \rangle - b_j : j = 1, \dots, m \}. \quad (3.4.7)$$

The steepest-descent trajectory can now be explicitly constructed, without calling for the machinery implied by Theorem 3.4.1: the key is that  $\hat{s}(\cdot)$  is “piecewise constant”, and the trajectory is made up of line-segments on successive time-intervals  $[\tau_k, \tau_{k+1}]$ ; see Fig. 2.2.4. This places us back in the framework of a minimization algorithm, with stepsizes  $t_k$  playing the role of time-differences  $\tau_{k+1} - \tau_k$ .

**Lemma 3.4.4** *For given  $x$  and  $d \neq 0$  in  $\mathbb{R}^n$ , define*

$$J^+ := \{j \notin J(x) : \langle s_j, d \rangle > f'(x, d)\},$$

*where  $J(x)$  is the active index-set of (2.1.1). For each  $j \in J^+$ , the equation in  $t$*

$$f(x) + tf'(x, d) = \langle s_j, x + td \rangle - b_j$$

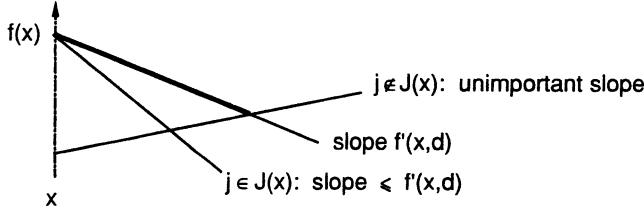
*has a positive solution, and call  $\bar{t} > 0$  the smallest of these solutions ( $\bar{t} = +\infty$  if  $J^+ = \emptyset$ ). Then, for all  $t \in [0, \bar{t}]$ ,*

$$f(x + td) = f(x) + tf'(x, d), \quad (3.4.8)$$

$$J(x + td) \subset J(x) \quad \text{hence} \quad \partial f(x + td) \subset \partial f(x). \quad (3.4.9)$$

PROOF. Just look at Fig. 3.4.1, drawn with a downhill  $d$ ;  $J^+$  appears as the “dangerous” index-set:  $j \notin J^+$  means

– either  $j \in J(x)$ ; then  $\langle s_j, d \rangle \leq f'(x, d)$  because a directional derivative is a max;



**Fig. 3.4.1.** Piecewise affine functions and “dangerous” indices

– or  $\langle s_j, d \rangle \leq f'(x, d)$ , from the very definition of  $J^+$ .

In both cases, the graph of  $f$  certainly does not meet the  $j^{\text{th}}$  affine piece in the direction  $d$ .

Finally, remember that subdifferentials are convex hulls of active gradients.  $\square$

This result is valid for any direction  $d$ . When  $d$  is actually the steepest-descent direction, we get something more:  $d$  remains steepest on the whole of  $[x, x + \bar{t}d]$ , but not at  $x + \bar{t}d$ .

**Lemma 3.4.5** *For given non-optimal  $x$ , take  $d = -\hat{s}(x)$  in Lemma 3.4.4. Then:*

- (i)  $\hat{s}(x + td) = \hat{s}(x)$  for all  $t \in [0, \bar{t}]$ ;
- (ii) assuming  $\bar{t} < +\infty$ , there holds  $\|\hat{s}(x + \bar{t}d)\| < \|\hat{s}(x)\|$ .

PROOF. [(i)] For all  $t \in [0, \bar{t}]$  and  $y \in \mathbb{R}^n$ , we can write

$$\begin{aligned} f(y) &\geq f(x) + \langle \hat{s}(x), y - x \rangle = && [\text{because } \hat{s}(x) \in \partial f(x)] \\ &= f(x) + \langle \hat{s}(x), y - x - td \rangle - t \|\hat{s}(x)\|^2 \\ &= f(x) + \langle \hat{s}(x), y - x - td \rangle + t f'(x, d) && [\text{Remark 1.3.7}] \\ &= f(x + td) + \langle \hat{s}(x), y - x - td \rangle; && [\text{because of (3.4.8)}] \end{aligned}$$

thus  $\hat{s}(x) \in \partial f(x + td)$ . Now we use Proposition 2.1.7:  $\hat{s}(x)$  satisfies

$$\langle s_j, \hat{s}(x) \rangle \geq \|\hat{s}(x)\|^2 \quad \text{for all } j \in J(x).$$

Because of (3.4.9), this inequality holds in particular for  $j \in J(x + \bar{t}d)$ .

In summary, we see that  $\hat{s}(x)$  satisfies the characterization of the orthogonal projection  $\hat{s}(x + \bar{t}d)$ .

[(ii)] Since the graph of  $\partial f$  is closed (Proposition VI.6.2.1), the property  $\hat{s}(x) \in \partial f(x + \bar{t}d)$  extends to  $t = \bar{t}$ , hence  $\|\hat{s}(x + \bar{t}d)\| \leq \|\hat{s}(x)\|$ . Equality holds if and only if  $\hat{s}(x + \bar{t}d) = \hat{s}(x)$ , but this is impossible: in Lemma 3.4.4, the definition of  $\bar{t}$  implies the existence of some index  $j \in J^+$ , hence

$$\langle s_j, \hat{s}(x) \rangle < \|\hat{s}(x)\|^2,$$

which is by construction in  $J(x + \bar{t}d)$ . In view of Proposition 2.1.7,  $\hat{s}(x)$  cannot be the projection of the origin onto  $\partial f(x + \bar{t}d)$ .  $\square$

The steepest-descent trajectory can now be constructed with the help of this result and Lemma 3.4.5(i).

**Algorithm 3.4.6 (Steepest-Descent Trajectory, Piecewise Affine Case)** The function to minimize is  $f$  of (3.4.7); the initial point  $x_1$  is given. Set  $k = 1$ .

STEP 1. Compute the active index-set  $J_k = J(x_k)$  and let  $\hat{s}_k$  solve the projection problem

$$\min \frac{1}{2} \|s\|^2 \quad \text{subject to } s \in \text{co} \{s_j : j \in J_k\}.$$

If  $\hat{s}_k = 0$  stop:  $x_k$  is optimal.

STEP 2. Set  $d_k = -\hat{s}_k$  and compute the stepsize  $t_k = \bar{t}$  as in Lemma 3.4.4. If  $\bar{t} = +\infty$  stop: the infimal value  $\bar{f}$  is  $-\infty$ .

STEP 3. Set  $x_{k+1} = x_k + t_k d_k$ , replace  $k$  by  $k + 1$  and loop to Step 1.  $\square$

**Remark 3.4.7** For the record, we mention again that a proper definition of  $J_k$  is numerically delicate (remember Remark 2.3.4). Our present algorithm actually belongs to the realm of linear algebra, and an incremental technique is convenient. Normally, the quadratic programming algorithm used in Step 1 computes the set  $J_1(x_k)$  alluded to in Remark 2.1.8. Then  $J_{k+1} = J_k(x_k) \cup \{j\}$ , where  $j$  is the index furnishing the stepsize  $\bar{t}$  from Lemma 3.4.4.

When using this technique,  $\bar{t}$  may become zero, hence the need for some other provision. We omit the details, which have little interest for our purpose.  $\square$

In the continuous case, convergence of the trajectory was established on the basis of (iii), (iv) from Theorem 3.4.1. The same arguments cannot be used here: the “abstract continuous time”  $\tau$  has nothing to do with the actual computing time spent by the algorithm. By contrast, Theorem 3.4.1(ii) seemed minor but it is now the key argument, via its discrete translation in Lemma 3.4.5(ii).

**Theorem 3.4.8** *Algorithm 3.4.6 stops after finitely many iterations.*

PROOF. The list of all possible active sets  $J_k$ ’s has at most  $2^m$  elements, and each such set characterizes its  $\hat{s}_k$ . In view of Lemma 3.4.5(ii), they are all different from each other; the algorithm stops after at most  $2^m$  iterations.  $\square$

Let us mention two interesting by-products:

- If a piecewise affine function is bounded from below, it has a minimum point; we knew it before (§V.3.4) but we have here a *constructive* and natural proof.
- If a piecewise affine function is unbounded from below, there are *fixed*  $x$  and  $d \neq 0$  such that

$$f(x + td) \downarrow -\infty \quad \text{when } t \rightarrow +\infty.$$

Actually, this last property holds for all  $x$  (and some  $d$ ), the reason being as follows: for all  $x$ , there are two constants  $\ell$  and  $L$  such that, for all  $d$  and  $t \geq 0$ ,

$$\ell + t \max_j \langle s_j, d \rangle \leq f(x + td) \leq L + t \max_j \langle s_j, d \rangle \tag{3.4.10}$$

(take for  $\ell$  and  $L$  respectively the min and max of  $\{\langle s_j, x \rangle - b_j\}$ ). Therefore, (3.4.10) is equivalent to  $\max_j \langle s_j, d \rangle < 0$ , a property independent of  $x$ .

### 3.5 Conclusion

The purpose of this Section 3 was mainly to show that the concept of steepest descent is not always the right one: even though it can be considered in some special situations (§3.4), sometimes it yields non-implementable algorithms (§3.1-2), sometimes it is irrelevant (§3.3) or intolerably slow (remember the very end of §II.2.2). Two more arguments may also be mentioned against it.

- First, it can be said that the whole idea is nothing but variations around known themes in ordinary nonlinear programming. When steepest-descent directions of a convex function can be computed in practice, it is usually because there is a formulation involving only smooth objective functions and constraints. See §2.3 and §3.1, including Remark 3.1.1. This does not imply that the approach is totally fruitless, though. It does have the merit of bringing new views on known fields.
- A second argument was alluded to in the beginning of §2: the user of any optimization algorithm must do his duty and provide the algorithm with the black box (U1) of Fig. II.1.2.1. In the present context, this duty is definitely more complicated than in classical optimization: the user must now compute the full subdifferential.

By contrast, all the optimization algorithms considered in the second volume of this book will request from the black box (U1) to compute *only* the value  $f(x)$  of the objective function and the value  $s(x)$  of a subgradient of  $f$  (an *arbitrary* subgradient) at every given  $x$ . Thus, we let  $s(x)$  denote the particular subgradient computed by the black box. The notation is slightly misleading, because it suggests that  $\partial f$  is always a singleton. It actually means that (U1) is assumed to be deterministic and answers two times the same  $s$  if called two times with the same  $x$ . Of course,  $s(\cdot)$  is a *discontinuous* mapping from  $\mathbb{R}^n$  to  $\mathbb{R}^n$ .

**Remark 3.5.1** This point was already seen in §3.1, more precisely in (3.1.3): denoting by  $s(x)$  the computed subgradient amounts to denoting by  $y(x)$  the optimal  $y$ . It implies that the operations involved during the  $y$ -maximization depend only on  $x$ , and not on the time of the day, say. A more concrete illustration is the following: consider the one-dimensional convex function

$$x \mapsto f(x) := \max \{0, x^2 - 1\}$$

which is kinky at  $x = \pm 1$ . For this  $f$ , the computation of  $s(x)$  in (U1) could be done as follows:

$$\text{if } |x| \leq 1 \text{ then } s = 0 \text{ else } s = 2x.$$

Another programmer might prefer

$$\text{if } |x| \geq 1 \text{ then } s = 2x \text{ else } s = 0.$$

Of course, the only possible variations concern the value of  $s$  at  $\pm 1$ , which are only bound to lie in  $[-2, 0]$  and  $[0, 2]$  respectively; but after (U1) is definitely loaded in the computer, we are entitled to call  $s(x)$  its output.

Actually, the difficulty from the point of view of minimizing this  $f$  is not that  $s(\pm 1)$  is ambiguously defined. The real difficulty is that  $s(\pm 1 + \varepsilon)$  has no relation

whatsoever with  $s(\pm 1 - \varepsilon)$ . In between, when  $x$  crosses  $\pm 1$ , some nasty event *must* happen (some call it a catastrophe). It is inherent to the nature of  $f$  and does happen no matter how (U1) is written, i.e. whichever choice is made for  $s(\pm 1)$ . Also, §3.3 suggests that a smoothed version of  $f$  may not be an efficient way of eliminating this nasty event.  $\square$

Let us sum up: from the point of view of the user, our minimization methods will present themselves just as “classical” methods, of the same type as in Chap. II. The user will have to compute  $f$  and  $s$ , without caring whether the latter is continuous. Very often, this makes his life much easier.

In this framework, our aim will be to develop minimization methods which perform reasonably well if  $s(\cdot)$  happens to be continuous, but which still converge in any case. Tables 3.3.2 and 3.3.3 are rather representative of what we have in mind.

## A. Appendix: Notations

We list in this appendix some basic concepts which are, or should be, well-known – but it is good sometimes to return to basics. This gives us the opportunity of making precise the system of notation used in this book. For example, some readers may have forgotten that “i.e.” means *id est*, the literal translation of “that is (to say)”. If we get closer to mathematics,  $S \setminus \{x\}$  denotes the set obtained by depriving a set  $S$  of a point  $x \in S$ . We also mention that, if  $f$  is a function,  $f^{-1}(y)$  is the *inverse image* of  $y$ , i.e. the set of all points  $x$  such that  $f(x) = y$ . When  $f$  is invertible, this set is the singleton  $\{f^{-1}(y)\}$ .

### 1 Some Facts About Optimization

**1.1** In the *totally ordered* set  $\mathbb{R}$ ,  $\inf E$  and  $\sup E$  are respectively the greatest lower bound – the *infimum* – and least upper bound – the *supremum* – of a nonempty subset  $E$ , when they exist (as real numbers). Then, they may or may not belong to  $E$ ; when they do, a more accurate notation is  $\min E$  and  $\max E$ . Whenever the relevant infima exist, the following relations are clear enough:

$$\left. \begin{aligned} \inf(E \cup F) &= \min \{\inf E, \inf F\}, \\ F \subset E &\implies \inf F \geq \inf E, \\ \inf(E \cap F) &\geq \max \{\inf E, \inf F\}. \end{aligned} \right\} \quad (1.1)$$

If  $E$  is characterized by a certain property  $P$ , we use the notation

$$E = \{r \in \mathbb{R} : r \text{ satisfies } P\}.$$

Defining (in  $\mathbb{R}$  considered as a real *vector space*) the standard operations on nonempty sets

$$\begin{aligned} E + F &:= \{r = e + f : e \in E, f \in F\}, \\ tE &:= \{tr : r \in E\} \quad \text{for } t \in \mathbb{R} \end{aligned}$$

(the sign “ $:=$ ” means “equals by definition”), it is also clear that

$$\left. \begin{aligned} \inf(E + F) &= \inf E + \inf F, \\ \inf tE &= t \inf E \quad \text{if } t > 0, \\ \inf(-E) &= -\sup E, \end{aligned} \right\} \quad (1.2)$$

whenever the relevant extrema exist.

The word *positive* means “ $> 0$ ”, and *nonpositive* therefore means “ $\leq 0$ ”; same conventions with negative and nonnegative. The set of nonnegative numbers is denoted by  $\mathbb{R}^+$  and, generally speaking, a substare deprives a set of the point 0. Thus, for example,

$$\mathbb{N}_* = \{1, 2, \dots\} \quad \text{and} \quad \mathbb{R}_*^+ = \{t \in \mathbb{R} : t > 0\}.$$

Squared brackets are used to denote the intervals of  $\mathbb{R}$ : for example,

$$\mathbb{R} \supset ]a, b] = \{t \in \mathbb{R} : a < t \leq b\}.$$

The symbol “ $\downarrow$ ” means convergence from the right, *the limit being excluded*; thus,  $t \downarrow 0$  means  $t \rightarrow 0$  in  $\mathbb{R}_*^+$ . The words “increasing” and “decreasing” are taken in a broad sense: a sequence  $\{t_k\}$  is increasing when  $k > k' \Rightarrow t_k \geq t_{k'}$ .

**1.2** Now, to denote a real-valued function  $f$  defined on a nonempty set  $X$ , we write

$$X \ni x \mapsto f(x) \in \mathbb{R}$$

and the *sublevel-set* of  $f$  at level  $r \in \mathbb{R}$  is defined by

$$S_r(f) := \{x \in X : f(x) \leq r\}.$$

If two functions  $f$  and  $g$  from  $X$  to  $\mathbb{R}$  satisfy

$$f(x) \leq g(x) \quad \text{for all } x \in X,$$

we say that  $f$  *minorizes*  $g$  (on  $X$ ), or that  $g$  *majorizes*  $f$ .

Computing the number

$$\inf \{f(x) : x \in X\} =: \bar{f} \tag{1.3}$$

represents a minimization problem posed in  $X$ : namely that of finding a so-called *minimizing sequence*, i.e.  $\{x_k\} \subset X$  such that  $f(x_k) \rightarrow \bar{f}$  when  $k \rightarrow +\infty$  (note that no structure is assumed on  $X$ ). In other words,  $\bar{f}$  is the largest lower bound of the subset  $f(X) \subset \mathbb{R}$ , and will often be called the infimal value, or more simply the *infimum* of  $f$  on  $X$ . Another notation for (1.3) is  $\inf_{x \in X} f(x)$ , or also  $\inf_X f$ . The function  $f$  is usually called the objective function, or also infimand. We can also meet supremands, minimands, etc.

From the relations (1.1), (1.2), we deduce (hereafter,  $\bar{f}_i$  denotes the infimum of  $f$  over  $X_i$  for  $i = 1, 2$ ):

$$\inf \{f(x) : x \in X_1 \cup X_2\} = \min \{\bar{f}_1, \bar{f}_2\},$$

$$X_1 \subset X_2 \implies \bar{f}_1 \geq \bar{f}_2,$$

$$\inf \{f(x) : x \in X_1 \cap X_2\} \geq \max \{\bar{f}_1, \bar{f}_2\},$$

$$\inf \{f(x_1) + f(x_2) : x_1 \in X_1 \text{ and } x_2 \in X_2\} = \bar{f}_1 + \bar{f}_2, \tag{1.4}$$

$$\inf \{tf(x) : x \in X\} = t\bar{f}, \quad \text{for } t \geq 0,$$

$$\inf \{-f(x) : x \in X\} = -\sup \{f(x) : x \in X\},$$

whenever the relevant extrema exist. The last relation is used very often.

The attention of the reader is drawn upon (1.4), perhaps the only non-totally trivial among the above relations. Calling  $E_1 := f(X_1)$  and  $E_2 := f(X_2)$  the *images* of  $X_1$  and  $X_2$  under  $f$ , (1.4) represents the sum of the infima  $\inf E_1$  and  $\inf E_2$ . There could just as well be two different infimands, i.e. (1.4) could be written more suggestively

$$\inf \{f(x_1) + g(x_2) : x_1 \in X_1 \text{ and } x_2 \in X_2\} = \bar{f}_1 + \bar{g}_2$$

( $g$  being another real-valued function). This last relation must not be confused with

$$\inf \{f(x) + g(x) : x \in X\} \geq \bar{f} + \bar{g};$$

here, in the language of (1.4),  $X_1 = X_2 = X$ , but only the image by  $f$  of the diagonal of  $X \times X$  is considered.

Another relation requiring some attention is the *decoupling*, or transitivity, of infima: if  $g$  sends the Cartesian product  $X \times Y$  to  $\mathbb{R}$ , then

$$\begin{aligned} & \inf \{g(x, y) : x \in X \text{ and } y \in Y\} = \\ & = \inf_{x \in X} [\inf_{y \in Y} g(x, y)] = \inf_{y \in Y} [\inf_{x \in X} g(x, y)]. \end{aligned} \tag{1.5}$$

**1.3** An *optimal solution* of (1.3) is an  $\bar{x} \in X$  such that

$$f(\bar{x}) = \bar{f} \leq f(x) \quad \text{for all } x \in X;$$

such an  $\bar{x}$  is often called a *minimizer*, a *minimum point*, or more simply a *minimum* of  $f$  on  $X$ . We will also speak of *global minimum*. To say that there exists a minimum is to say that the inf in (1.3) is a min; the infimum  $\bar{f} = f(\bar{x})$  can then be called the *minimal value*. The notation

$$\min \{f(x) : x \in X\}$$

is the same as (1.3), and says that there does exist a solution; we stress the fact that both notations represent at the same time a *number* and a *problem to solve*. It is sometimes convenient to denote by

$$\operatorname{Argmin} \{f(x) : x \in X\}$$

the set of optimal solutions of (1.3), and to use “argmin” if the solution is unique.

It is worth mentioning that the decoupling property (1.5) has a translation in terms of Argmin’s. More precisely, the following properties are easy to see:

- If  $(\bar{x}, \bar{y})$  minimizes  $g$  over  $X \times Y$ , then  $\bar{y}$  minimizes  $g(\bar{x}, \cdot)$  over  $Y$  and  $\bar{x}$  minimizes over  $X$  the function

$$\varphi(x) := \inf \{g(x, y) : y \in Y\}.$$

- Conversely, if  $\bar{x}$  minimizes  $\varphi$  over  $X$  and if  $\bar{y}$  minimizes  $g(\bar{x}, \cdot)$  over  $Y$ , then  $(\bar{x}, \bar{y})$  minimizes  $g$  over  $X \times Y$ .

Needless to say, symmetric properties are established, interchanging the roles of  $x$  and  $y$ .

**1.4** In our context,  $X$  is equipped with a topology; actually  $X$  is a subset of some finite-dimensional real vector space, call it  $\mathbb{R}^n$ ; the topology is then that induced by a norm. The *interior* and *closure* of  $X$  are denoted  $\text{int } X$  and  $\text{cl } X$  respectively.

The concept of limit is assumed familiar. We recall that the *limes inferior* (in the ordered set  $\mathbb{R}$ ) is the smallest cluster point.

**Remark 1.1** The standard terminology is *lower limit* (“abbreviated” as  $\liminf$ !). This terminology is unfortunate, however: a limit must be a well-defined unique element; otherwise, expressions such as “ $f(x)$  has a limit” are ambiguous.  $\square$

Thus, to say that  $\ell = \liminf_{x \rightarrow x^*} f(x)$ , with  $x^* \in \text{cl } X$ , means: for all  $\varepsilon > 0$ ,

there is a neighborhood  $N(x^*)$  such that  $f(x) \geq \ell - \varepsilon$  for all  $x \in N(x^*)$   
and

in any neighborhood  $N(x^*)$ , there is an  $x \in N(x^*)$  such that  $f(x) \leq \ell + \varepsilon$ ;

in particular, if  $x^* \in X$ , we certainly have  $\ell \leq f(x^*)$ .

Let  $x^* \in X$ . If  $f(x^*) \leq \liminf_{x \rightarrow x^*} f(x)$ ,  $f$  is said to be *lower semi-continuous* (l.s.c) at  $x^*$ ; and *upper semi-continuous* when  $f(x^*) \geq \limsup_{x \rightarrow x^*} f(x)$ . It is well-known that, if  $X$  is a compact set on which  $f$  is continuous, then the lower bound  $\bar{f}$  exists and (1.3) has a solution. Actually, lower semi-continuity (of  $f$  on the whole compact  $X$ ) suffices: if  $\{x_k\}$  is a minimizing sequence, with some cluster point  $x^* \in X$ , we have

$$f(x^*) \leq \liminf_{k \rightarrow \infty} f(x_k) = \lim_{k \rightarrow \infty} f(x_k) = \bar{f}.$$

Another observation is: let  $E$  be such that  $\text{cl } E \subset X$ ; if  $f$  is continuous on  $\text{cl } E$ , then

$$\inf \{f(x) : x \in E\} = \inf \{f(x) : x \in \text{cl } E\}.$$

This relation is wrong if  $f$  is only l.s.c, though: then, only (1.1) gives useful relations.

Related with (1.3), another problem is whether a given minimizing sequence  $\{x_k\}$  converges to an optimal solution when  $k \rightarrow +\infty$ . This problem is really distinct from (1.3): for example, with

$$X := \mathbb{R}, \quad f(0) := 0, \quad f(x) := 1/|x| \text{ for } x \neq 0$$

the sequence defined by  $x_k = k$  is minimizing but does not converge to the minimum 0 when  $k \rightarrow +\infty$ .

## 2 The Set of Extended Real Numbers

In convex analysis and optimization, there are serious reasons to give a meaning to (1.3), for arbitrary  $f$  and  $X$ . For this, two additional elements are appended to  $\mathbb{R}$ :  $+\infty$  and  $-\infty$ .

If  $E \subset \mathbb{R}$  is nonempty but unbounded from above, we set  $\sup E = +\infty$ ; similarly,  $\inf E = -\infty$  if  $E$  is unbounded from below. Then consider the case of an empty set: to maintain a relation such as (1.1)

$$\inf(E \cup \emptyset) [= \inf E] = \min \{\inf E, \inf \emptyset\} \quad \text{for all } \emptyset \neq E \subset \mathbb{R},$$

we have no choice and we set

$$\inf \emptyset = +\infty;$$

naturally,  $\sup \emptyset = -\infty$ , and this maintains the relation  $\inf(-E) = -\sup E$  in (1.2).

It should be noted that the world of convex analysis and minimization, which starts at (1.3), is not symmetric:  $+\infty$  and  $-\infty$  do not play the same role, and it suffices for our purpose to consider the set  $\mathbb{R} \cup \{+\infty\}$ . Extending the notation of the intervals of  $\mathbb{R}$ , this set will also be denoted by  $] -\infty, +\infty ]$ .

To extend the structure of  $\mathbb{R}$  to this new set, the natural rules are adopted:

- order:  $x \leqslant +\infty$  for all  $x \in \mathbb{R} \cup \{+\infty\}$ ;
- addition:  $(+\infty) + x = x + (+\infty) = +\infty$  for all  $x \in \mathbb{R} \cup \{+\infty\}$ ;
- multiplication:  $t \cdot (+\infty) = +\infty$  for all  $0 < t \in \mathbb{R} \cup \{+\infty\}$ .

Thus, we see that

- the structured set  $(\mathbb{R} \cup \{+\infty\}, +)$  is not a group, just because  $+\infty$  has no opposite;
- it is a fortiori not a field, a second reason being that we avoid writing  $t \times (+\infty)$  for  $t \leqslant 0$ .

On the other hand, we leave it to the reader to check that the other axioms are preserved (for the order, the addition and the multiplication); so some calculus can at least be done in  $\mathbb{R} \cup \{+\infty\}$ .

Actually,  $\mathbb{R} \cup \{+\infty\}$  is nothing more than an *ordered convex cone*, analogous to the set  $\mathbb{R}_*^+$  of positive numbers. In particular, observe the following continuity properties:

$$(x_k, y_k) \rightarrow (x, y) \text{ in } [\mathbb{R} \cup \{+\infty\}]^2 \implies x_k + y_k \rightarrow x + y \text{ in } \mathbb{R} \cup \{+\infty\}; \\ (t_k, x_k) \rightarrow (t, x) \text{ in } \mathbb{R}_*^+ \times [\mathbb{R} \cup \{+\infty\}] \implies t_k x_k \rightarrow tx \text{ in } \mathbb{R} \cup \{+\infty\}.$$

In this book, the minimization problems of §1 – and in particular (1.3) – will be understood as posed in  $\mathbb{R} \cup \{+\infty\}$ . The advantage of this is to give a systematic meaning to all the relations of §1. On the other hand, the reader should not feel too encumbered by this new set, which takes the place of the familiar set of real numbers where algebra is “easy”. First of all,  $\mathbb{R} \cup \{+\infty\}$  is relevant only as far as images of functions are concerned: any algebraic manipulations involving no term  $f(x)$  is “safe” and requires no special attention. When some  $f(x)$  is involved, the following pragmatic attitude can be adopted:

- comparison and addition: no problems in  $\mathbb{R} \cup \{+\infty\}$ , just as in  $\mathbb{R}$ ;
- subtraction: before subtracting  $f(x)$ , make sure that  $f(x) < +\infty$ ;
- multiplication: think of a term like  $t f(x)$  as the multiplication of the vector  $f(x)$  by the scalar  $t$ ; if  $t \leqslant 0$ , make sure that  $f(x) < +\infty$  (note: in convex analysis and optimization, the product of functions  $f(x)g(x)$  is rarely used, and multiplication by  $-1$  puts (1.3) in a different world);
- division: same problems as in  $\mathbb{R}$ , namely avoid division by 0;
- convergence: same problems as in  $\mathbb{R}$ , namely pay attention to  $\infty - \infty$  and  $0 \cdot (+\infty)$ ;
- in general, do not abuse expressions like  $t f(x)$  with  $t \leqslant 0$ , or  $r - f(x)$ , etc.: they do not fit well with the conical structure of  $\mathbb{R} \cup \{+\infty\}$ .

### 3 Linear and Bilinear Algebra

**3.0** Let us start with the *model-situation* of  $\mathbb{R}^n$ , the real  $n$ -dimensional vector space of  $n$ -uples  $x = (\xi^1, \dots, \xi^n)$ . In this space, the vectors  $e_1, \dots, e_n$ , where each  $e_i$  has coordinates  $(0, \dots, 0, 1, 0, \dots, 0)$  (the “1” in  $i^{\text{th}}$  position) form a basis, called the *canonical basis*. The linear mappings from  $\mathbb{R}^m$  to  $\mathbb{R}^n$  are identified with the  $n \times m$  *matrices* which represent them in the canonical bases; *vectors* of  $\mathbb{R}^n$  are thus naturally identified with  $n \times 1$  matrices.

The space  $\mathbb{R}^n$  is equipped with the canonical, or standard, Euclidean structure with the help of the scalar product

$$x = (\xi^1, \dots, \xi^n), \quad y = (\eta^1, \dots, \eta^n) \quad \mapsto \quad x^\top y := \sum_{i=1}^n \xi^i \eta^i$$

(also denoted by  $x \cdot y$ ). Then we can speak of the Euclidean space  $(\mathbb{R}^n, \cdot^\top \cdot)$ .

**3.1** More generally, a *Euclidean space* is a real vector space, say  $X$ , of *finite dimension*, say  $n$ , equipped with a *scalar product* denoted by  $\langle \cdot, \cdot \rangle$ . Recall that a scalar (or inner) product is a bilinear symmetric mapping  $\langle \cdot, \cdot \rangle$  from  $X \times X$  to  $\mathbb{R}$ , satisfying  $\langle x, x \rangle > 0$  for  $x \neq 0$ .

(a) If a basis  $\{b_1, \dots, b_n\}$  has been chosen in  $X$ , along which two vectors  $x$  and  $y$  have the coordinates  $(\xi^1, \dots, \xi^n)$  and  $(\eta^1, \dots, \eta^n)$ , we have

$$\langle x, y \rangle = \sum_{i,j=1}^n \xi^i \eta^j \langle b_i, b_j \rangle.$$

This can be written  $\langle x, y \rangle = x^\top Q y$ , where  $Q$  is a symmetric positive definite  $n \times n$  matrix. In this situation, to equip  $X$  with a scalar product is actually to take a symmetric positive definite matrix.

The simplest matrix  $Q$  is the identity matrix  $I$ , or  $I_n$ , which corresponds to the scalar product

$$\langle x, y \rangle = x^\top y = \sum_{i=1}^n \xi^i \eta^i,$$

called the *dot-product*. For this particular product, one has  $\langle b_i, b_j \rangle = \delta_{ij}$  ( $\delta_{ij}$  is the symbol of Kronecker:  $\delta_{ij} = 0$  if  $i \neq j$ ,  $\delta_{ii} = 1$ ). The basis  $\{b_1, \dots, b_n\}$  is said *orthonormal* for this scalar product; and this scalar product is of course the only one for which the given basis is orthonormal.

Thus, whenever we have a basis in  $X$ , we know all the possible ways of equipping  $X$  with a Euclidean structure.

(b) Reasoning the other direction, let us start from a Euclidean space  $(X, \langle \cdot, \cdot \rangle)$  of dimension  $n$ . It is possible to find a basis  $\{b_1, \dots, b_n\}$  of  $X$ , which is orthonormal for

the given scalar product (i.e. which satisfies  $\langle b_i, b_j \rangle = \delta_{ij}$  for  $i, j = 1, \dots, n$ ). If two vectors  $x$  and  $y$  are expressed in terms of this basis,  $\langle x, y \rangle$  can be written  $x^\top y$ .

Use the space  $\mathbb{R}^n$  of §3.0 and denote by  $\varphi : \mathbb{R}^n \rightarrow X$  the unique *linear mapping* (isomorphism of vector spaces) satisfying  $\varphi(e_i) = b_i$  for  $i = 1, \dots, n$ . Then

$$x^\top y = \langle \varphi(x), \varphi(y) \rangle \quad \text{for all } x \text{ and } y \text{ in } \mathbb{R}^n,$$

so the Euclidean structure is also carried over by  $\varphi$ , which is therefore an isomorphism of Euclidean spaces as well. Thus, *any* Euclidean space  $(X, \langle \cdot, \cdot \rangle)$  of dimension  $n$  is *isomorphic* to  $(\mathbb{R}^n, \top)$ , which explains the importance of this last space. However, given a Euclidean space, an orthonormal basis need not be easy to construct; said otherwise, one must sometimes content oneself with a scalar product imposed by the problem considered.

**Examples 3.1** (i) An important space for applications is the Sobolev space  $H^1(\Omega)$ , with for example  $\Omega = ]0, 1[$ , in which the scalar product of two functions  $x$  and  $y$  is

$$\int_0^1 [x(t)y(t) + \dot{x}(t)\dot{y}(t)]dt. \quad (3.1)$$

For a simple discretization of this function space, set  $h = 1/n$  and, for  $i = 1, \dots, n$ , let  $\xi^i, \eta^i$  approximate the mean value of  $x, y$  on  $]i-1)h, ih[$ ; we have now a vector space of  $n$ -uples (note, however, that more sophisticated discretizations can be used). To take into account the derivatives in (3.1), append to any vector  $(\xi^1, \dots, \xi^n) \in \mathbb{R}^n$  the two dummy coordinates  $\xi^0 = \xi^{n+1} = 0$ . Then the following scalar product is natural:

$$h \sum_{i=1}^n \xi^i \eta^i + h \sum_{i=0}^n \frac{\xi^{i+1} - \xi^i}{h} \frac{\eta^{i+1} - \eta^i}{h}.$$

(ii) Vector spaces of matrices form a rich field of applications for the techniques and results of convex analysis and optimization. The set of  $p \times q$  matrices forms a vector space of dimension  $p + q$ , in which the scalar product of two matrices  $M$  and  $N$  defined by

$$\langle M, N \rangle := \text{tr } M^\top N \quad [= \text{trace of } M^\top N]$$

is a natural one. □

(c) A subspace  $V$  of  $(X, \langle \cdot, \cdot \rangle)$  can be equipped with the Euclidean structure defined by

$$V \times V \ni (x, y) \mapsto \langle x, y \rangle.$$

Unless otherwise specified, we will generally use this *induced* structure, with the same notation for the scalar product in  $V$  and in  $X$ .

More importantly, let  $(X_1, \langle \cdot, \cdot \rangle_1)$  and  $(X_2, \langle \cdot, \cdot \rangle_2)$  be two Euclidean spaces. Their Cartesian product  $X = X_1 \times X_2$  can be made Euclidean via the scalar product

$$((x_1, x_2), (y_1, y_2)) = (x, y) \mapsto \langle x, y \rangle := \langle x_1, y_1 \rangle_1 + \langle x_2, y_2 \rangle_2.$$

This is not compulsory: cases may occur in which the product-space  $X$  has its own Euclidean structure, not possessing this “decomposability” property.

**3.2** Let  $(X, \langle \cdot, \cdot \rangle)$  and  $(Y, \langle\langle \cdot, \cdot \rangle\rangle)$  be two Euclidean spaces, knowing that we could write just as well  $X = \mathbb{R}^n$  and  $Y = \mathbb{R}^m$ .

(a) If  $A$  is a linear operator from  $X$  to  $Y$ , the *adjoint* of  $A$  is the unique operator  $A^*$  from  $Y$  to  $X$ , defined by

$$\langle A^*y, x \rangle = \langle\langle y, Ax \rangle\rangle \quad \text{for all } (x, y) \in X \times Y.$$

There holds  $(A^*)^* = A$ . When both  $X$  and  $Y$  have orthonormal bases (as is the case with canonical bases for the dot-product in the respective spaces), the matrix representing  $A^*$  in these bases is the *transpose* of the matrix representing  $A$ .

Consider the case  $(Y, \langle\langle \cdot, \cdot \rangle\rangle) = (X, \langle \cdot, \cdot \rangle)$ . When  $A$  is invertible, so is  $A^*$ , and  $(A^*)^{-1} = (A^{-1})^*$ . When  $A^* = A$ , we say that  $A$  is *self-adjoint*, or *symmetric*. If, in addition,

$$\langle Ax, x \rangle > 0 \quad [\text{resp. } \geq 0] \quad \text{for all } 0 \neq x \in X,$$

then  $A$  is *positive definite* [resp. *positive semi-definite*]. When  $X = Y$  is equipped with an orthonormal basis, symmetric operators can be characterized in terms of matrices:  $A$  is symmetric [resp. symmetric positive (semi)-definite] if and only if the matrix representing  $A$  (in the orthonormal basis) is symmetric [resp. symmetric positive (semi)-definite].

(b) When the image-space  $Y$  is  $\mathbb{R}$ , an operator is rather called a *form*. If  $\ell$  is a linear form on  $(X, \langle \cdot, \cdot \rangle)$ , there exists a unique  $s \in X$  such that  $\ell(x) = \langle s, x \rangle$  for all  $x \in X$ . If  $q$  is a quadratic form on  $(X, \langle \cdot, \cdot \rangle)$ , there exists a unique symmetric operator  $Q$  such that

$$q(x) := \frac{1}{2}\langle Qx, x \rangle \quad \text{for all } x \in X$$

(the coefficient 1/2 is useful to simplify most algebraic manipulations).

**Remark 3.2** The correspondence  $\ell \leftrightharpoons s$  is a triviality in  $(\mathbb{R}^n, \top)$  (just transpose the  $1 \times n$  matrices to vectors) but this is deceiving. Indeed, it is the correspondence  $X \leftrightharpoons X^*$  between a space and its *dual* that is being considered. For two vectors  $s$  and  $x$  of  $X$ , it is good practice to think of the scalar product  $\langle s, x \rangle$  as the action of the first argument  $s$  (a slope, representing an element in the dual) on the second argument  $x$ ; this helps one to understand what one is doing. Likewise, the operator  $Q$  associated with a quadratic form sends  $X$  to  $X^*$ ; and an adjoint  $A^*$  is from  $Y^*$  to  $X^*$ .  $\square$

**3.3** Two subspaces  $U$  and  $V$  of  $(X, \langle \cdot, \cdot \rangle)$  are mutually *orthogonal* if

$$\langle u, v \rangle = 0 \quad \text{for all } u \in U \text{ and } v \in V,$$

a relation denoted by  $U \perp V$ . On the other hand,  $U$  and  $V$  are *generators* of  $X$  if  $U + V = X$ . For given  $U$ , we denote by  $U^\perp$  the *orthogonal supplement* of  $U$ , i.e. the unique subspace orthogonal to  $U$  such that  $U$  and  $U^\perp$  form a generator of  $X$ .

Let  $A : X \rightarrow Y$  be an arbitrary linear operator,  $X$  and  $Y$  having arbitrary scalar products. As can easily be seen,

$$\text{Ker } A := \{x \in X : Ax = 0 \in Y\}$$

and

$$\text{Im } A^* := \{x \in X : x = A^*s \text{ for some } s \in Y\}$$

are orthogonal generators of  $X$ . In other words,

$$\text{Ker } A = (\text{Im } A^*)^\perp.$$

This is a very important relation; one must learn to use it quasi-mechanically, remembering that  $A^{**} = A$  and  $U^{\perp\perp} = U$ . For example, if  $A$  is symmetric,  $\text{Im } A$  is the orthogonal supplement of  $\text{Ker } A$ .

Important examples of linear operators from  $(X, \langle \cdot, \cdot \rangle)$  to itself are *orthogonal projections*: if  $H$  is a subspace of  $X$ , the operator  $p_H : X \rightarrow X$  of orthogonal projection onto  $H$  is defined by:

$$\begin{cases} p_H x = 0 & \text{for } x \in H^\perp, \\ p_H x = x & \text{for } x \in H, \\ p_H \text{ is completed by linearity in between.} \end{cases}$$

This  $p_H$  is *symmetric* and *idempotent* (i.e.  $p_H \circ p_H = p_H$ ). Conversely, a linear operator  $p$  which is symmetric and idempotent is an orthogonal projection; of course, it is the projection onto the subspace  $\text{Im } p$ .

**3.4** If  $A$  is a symmetric linear operator on  $X$ , remember that  $(\text{Im } A)^\perp = \text{Ker } A$ . Then consider the operator  $p_{\text{Im } A}$  of orthogonal projection onto  $\text{Im } A$ . For given  $y \in X$ , there is a unique  $x = x(y)$  in  $\text{Im } A$  such that  $Ax = p_{\text{Im } A}y$ ; furthermore, the mapping  $y \mapsto x(y)$  is linear. This mapping is called the *pseudo-inverse*, or generalized inverse, of  $A$  (more specifically, it the pseudo-inverse of Moore and Penrose). We denote it by  $A^-$ ; other notations are  $A^+$ ,  $A^\#$ , etc.

We recall some useful properties of the pseudo-inverse:  $\text{Im } A^- = \text{Im } A$ ;  $A^- A = AA^- = p_{\text{Im } A}$ ; and if  $A$  is positive semi-definite, so is  $A^-$ .

## 4 Differentiation in a Euclidean Space

A Euclidean space  $(X, \langle \cdot, \cdot \rangle)$  is a normed vector space (certainly complete) thanks to the norm

$$X \ni x \mapsto \|x\| := \sqrt{\langle x, x \rangle},$$

called the *Euclidean norm* associated with  $\langle \cdot, \cdot \rangle$ . We denote by

$$B(x, r) := \{y \in X : \|y - x\| \leq r\}$$

the *ball* of center  $x \in X$  and radius  $r > 0$ . In particular,  $B(0, 1)$  is called the unit ball, whose boundary is the *unit sphere*

$$\tilde{B}(0, 1) := \{y \in X : \|y\| = 1\}.$$

The norm and scalar product are related by the fundamental *Cauchy-Schwarz inequality* (no “t”, please)

$$|\langle s, x \rangle| \leq \|s\| \|x\| \quad \text{for all } (s, x) \in X \times X.$$

Remember that all norms are equivalent in our finite-dimensional space  $X$ : if  $\|\cdot\|$  is another norm, there are positive numbers  $\ell$  and  $L$  such that

$$\ell\|x\| \leq \|x\| \leq L\|x\| \quad \text{for all } x \in X.$$

However, the Euclidean norm  $\|\cdot\|$  plays a special role.

**4.1** We will denote by  $\varepsilon(\cdot)$  a generic function (from some normed space to another) which tends to 0 when its argument tends to 0. For example, the continuity of  $f$  at  $x$  can be expressed by

$$f(x + h) = f(x) + \varepsilon(h).$$

When we need to distinguish *speeds* of convergence, multiplicative factors can be used. For example,

$$\|h\|^\alpha \varepsilon(h) \quad \text{for } \alpha = 1, 2, \dots$$

denotes the functions tending to 0 faster than  $\|h\|$ ,  $\|h\|^2$ , ... A more handy notation for  $\|h\|^\alpha \varepsilon(h)$  is  $o(\|h\|^\alpha)$  (pronounce “little oh of ...”).

Beware that these are only notations, and algebraic manipulation with them should be done very carefully. For example  $\varepsilon(\cdot) \equiv r\varepsilon(\cdot)$  for all  $r \neq 0$ , hence in particular  $\varepsilon(\cdot) - \varepsilon(\cdot) = \varepsilon(\cdot)$ ! Always keep the definitions in mind; for example, to say that a function  $h \mapsto \varphi(h)$  is  $o(\|h\|)$  means:

$$\forall \varepsilon > 0 \quad \exists \delta > 0 \quad \text{such that} \quad \|h\| \leq \delta \implies \|\varphi(h)\| \leq \varepsilon \|h\|.$$

With this last notation, a function  $f : \Omega \rightarrow \mathbb{R}$ , defined on an open set  $\Omega \subset X$ , is said to be *differentiable* at  $x \in \Omega$  if there exists a linear form  $\ell$  on  $X$  such that

$$f(x + h) = f(x) + \ell(h) + o(\|h\|).$$

This linear form  $\ell$ , denoted by  $f'(x)$ ,  $Df(x)$  or  $df(x)$ , is called the *differential* of  $f$  at  $x$ . According to §3.2(b), it can be represented by a unique element of  $X$ ; this element is called the *gradient* of  $f$  at  $x$ , denoted by  $\nabla f(x)$ , and is therefore defined by

$$f'(x)(h) = \langle \nabla f(x), h \rangle \quad \text{for all } h \in X.$$

**Example 4.1** Let  $H \subset X$  be a subspace, equipped with the Euclidean structure induced by  $(X, \langle \cdot, \cdot \rangle)$  as in §3.1(c). If  $f$  is differentiable at  $x \in H$ , its gradient  $\nabla f(x)$  is obtained from

$$f(x + h) = f(x) + \langle \nabla f(x), h \rangle + o(\|h\|). \tag{4.1}$$

Then define the function  $f_H : H \rightarrow \mathbb{R}$  to be the restriction of  $f$  to  $H$ . This  $f_H$  is differentiable at  $x$  and its gradient  $\nabla f_H(x)$  is the vector (of  $H$ !) satisfying

$$f(x + h) = f(x) + \langle \nabla f_H(x), h \rangle + o(\|h\|) \quad \text{for all } h \in H.$$

Its computation is simple: in view of the properties of an orthogonal projection,

$$\langle \nabla f(x), h \rangle = \langle p_H \nabla f(x), h \rangle \quad \text{for all } h \in H;$$

Plugging this into (4.1), we see that  $\nabla f_H(x) = p_H \nabla f(x)$ .  $\square$

It is important to realize that the representation of  $f'(x)$  by  $\nabla f(x)$  changes if  $\langle \cdot, \cdot \rangle$  is changed. The gradient depends on the scalar product; but the differential does not.

If the space is equipped with an orthonormal basis, along which  $x$  has the coordinates  $\xi^1, \dots, \xi^n$ , then  $f'(x)$  is represented by the *row-matrix*

$$\left[ \frac{\partial f}{\partial \xi^1}(x), \dots, \frac{\partial f}{\partial \xi^n}(x) \right]$$

and  $\nabla f(x)$  is the *vector* of  $\mathbb{R}^n$  whose coordinates are  $(\partial f / \partial \xi^i)(x)$  for  $i = 1, \dots, n$ .

**4.2** More generally, a function  $F$  from  $\Omega \subset X$  to some other Euclidean space, say  $Y = \mathbb{R}^m$ , is differentiable at  $x \in \Omega$  if there exists a linear operator  $L$  from  $X$  to  $Y$  such that

$$F(x + h) = F(x) + L(h) + o(\|h\|).$$

The differential  $L$  of  $F$  at  $x$  is also called the *Jacobian operator* of  $F$  at  $x$ , again denoted by  $F'(x)$ ,  $DF(x)$ , or also  $JF(x)$ . Nothing is really new with respect to the scalar case of §4.1; denoting by  $f_1, \dots, f_m$  the component-functions of  $F$  along some basis of  $Y$ ,  $F$  is differentiable at  $x$  if and only if each  $f_j$  is such and

$$JF(x)(h) = (f'_1(x)(h), \dots, f'_m(x)(h)) \quad \text{for all } h \in X.$$

The matrix representation of  $JF(x)$  along the bases of  $X$  and  $Y$  is an  $m \times n$  matrix, whose  $(i, j)^{\text{th}}$  element is  $(\partial f_i / \partial \xi^j)(x)$ .

Given a scalar-valued function  $f$ , differentiable on  $\Omega$ , consider the function  $y \mapsto f'(y)$ , sending  $\Omega$  to the space of linear forms on  $X$ . If this new function is in turn differentiable at  $x$ , we obtain the *second-order* differential (of  $f$  at  $x$ ). This defines a *bilinear form* via

$$X \times X \ni (h, k) \mapsto [(f')'(x)(h)](k) =: f''(x)(h, k),$$

which is also *symmetric*; as such, it induces a quadratic form on  $X$  (for which we will use the same notation).

If  $X$  is equipped with a scalar product  $\langle \cdot, \cdot \rangle$ , §3.2(b) tells us that the quadratic form  $f''(x)$  defines a symmetric operator: the *Hessian* of  $f$  at  $x$ , denoted by  $\nabla^2 f(x)$ , or  $Hf(x)$ . Just as the gradient, the Hessian depends on the scalar product; and there holds the *second-order approximation* of  $f$  at  $x$ :

$$f(x + h) = f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} \langle \nabla^2 f(x)h, h \rangle + o(\|h\|^2).$$

With an orthonormal basis and  $x = (\xi^1, \dots, \xi^n)$ ,  $\nabla^2 f(x)$  is represented by a symmetric matrix whose  $(i, j)^{\text{th}}$  element is  $(\partial^2 f / \partial \xi^i \partial \xi^j)(x)$ , called the *Hessian matrix* of  $f$  at  $x$ .

**4.3** For an illustration, consider the space  $X$  of  $n \times n$  matrices, equipped with the scalar product  $\langle\langle M, N \rangle\rangle = \text{tr}(M^\top N)$  of Example 3.1(ii). On this space, take the determinant-function:

$$X \ni M \mapsto f(M) = \det M.$$

Develop  $\det M$  along the  $i^{\text{th}}$  column of  $M = [m_{ij}]$  to obtain

$$\det M = m_{i1}c_{i1} + m_{i2}c_{i2} + \cdots + m_{in}c_{in},$$

$c_{ij}$  denoting the  $(i, j)^{\text{th}}$  cofactor of  $M$ . Thus, calling  $\text{cof } M = [c_{ij}]$  the matrix of cofactors of  $M$ , we have

$$\frac{\partial \det}{\partial m_{ij}}(M) = c_{ij} \quad \text{for } i, j = 1, \dots, n,$$

hence

$$X \ni P = [p_{ij}] \mapsto (\det)'(M)(P) = \sum_{i,j=1}^n c_{ij} p_{ij} = \text{tr}((\text{cof } M)^\top P).$$

This shows that  $\nabla(\det)(M) = \text{cof } M$  [ $= (\det M)[M^{-1}]^\top$  if  $M$  is invertible].

Now, with  $K^+$  denoting the set of symmetric positive definite  $n \times n$  matrices (an open set in  $X$ ), consider the function

$$K^+ \ni M \mapsto \log(\det M).$$

This new function is of interest because maximizing the determinant in a subset of  $K^+$  is equivalent to maximizing its logarithm, and cannot produce a singular matrix:  $\log \det$  acts as a *barrier* in  $K^+$ . Using standard calculus rules, we obtain

$$\nabla(\log \det)(M) = M^{-1};$$

to memorize this formula, think of the derivative of  $\log$  on  $\mathbb{R}_*^+$ !

## 5 Set-Valued Analysis

**5.1** If  $S$  is a nonempty closed subset of  $\mathbb{R}^n$ , we denote by

$$d_S(x) := \min_{y \in S} \|y - x\|$$

the *distance* from  $x$  to  $S$ . Given two nonempty closed sets  $S_1$  and  $S_2$ , consider

$$e_H(S_1/S_2) := \sup \{d_{S_2}(x) : x \in S_1\},$$

called the *excess* of  $S_1$  over  $S_2$ : geometrically,

$$e_H(S_1/S_2) \leq \varepsilon \quad \text{means} \quad S_1 \subset S_2 + B(0, \varepsilon).$$

The *Hausdorff-distance*  $\Delta_H$  between  $S_1$  and  $S_2$  is then the symmetrization of the above concept:

$$\Delta_H(S_1, S_2) := \max \{e_H(S_1/S_2), e_H(S_2/S_1)\}.$$

One checks immediately that  $\Delta_H(S_1, S_2) \in \mathbb{R}^+ \cup \{+\infty\}$ , but  $\Delta_H$  is a finite-valued function when restricted to *bounded* closed sets. Also

$$\begin{aligned}\Delta_H(S_1, S_2) = 0 &\iff S_1 = S_2, \\ \Delta_H(S_1, S_2) &= \Delta_H(S_2, S_1), \\ \Delta_H(S_1, S_3) &\leq \Delta_H(S_1, S_2) + \Delta_H(S_2, S_3).\end{aligned}$$

In other words,  $\Delta_H$  does define a distance on the family of nonempty compact subsets of  $\mathbb{R}^n$ .

**5.2** A mapping  $F$  which, to  $x \in X$ , associates a subset of  $\mathbb{R}^n$ , is called a multi-valued, or set-valued mapping, or more simply a *multipunction*; we use the notation

$$X \ni x \longmapsto F(x) \subset \mathbb{R}^n.$$

The *domain*  $\text{dom } F$  of  $F$  is the set of  $x \in X$  such that  $F(x) \neq \emptyset$ . Its *image* (or range)  $F(X)$  and *graph*  $\text{gr } F$  are the unions of the sets  $F(x) \subset \mathbb{R}^n$  and  $\{x\} \times F(x) \subset X \times \mathbb{R}^n$  respectively, when  $x$  describes  $X$  (or, more precisely,  $\text{dom } F$ ). A *selection* of  $F$  is a particular function  $f : \text{dom } F \rightarrow \mathbb{R}^n$  with  $f(x) \in F(x)$  for all  $x$ .

The concept of convergence is here much more tricky than in the single-valued case. First of all, since a limit is going to be a set anyway, the following concept is relevant: the *limes exterior* of  $F(x)$  for  $x \rightarrow x^*$  is the set of all cluster points of all selections (here,  $x^* \in \text{cl dom } F$ ). In other words,  $y \in \lim \text{ext}_{x \rightarrow x^*} F(x)$  means: there exists a sequence  $\{x_k, y_k\}_k$  such that

$$y_k \in F(x_k), \quad x_k \rightarrow x^* \text{ and } y_k \rightarrow y \quad \text{when } k \rightarrow +\infty.$$

Note that this does not depend on multi-valuedness: each  $F(x)$  might well be a singleton for all  $x$ . For example,

$$\lim_{t \downarrow 0} \text{ext}\{\sin(1/t)\} = [-1, +1].$$

The *limes interior* of  $F(x)$  for  $x \rightarrow x^*$  is the set of limits of all convergent selections:  $y \in \lim \text{int}_{x \rightarrow x^*} F(x)$  means that there exists a function  $x \mapsto f(x)$  such that

$$f(x) \in F(x) \text{ for all } x \quad \text{and} \quad f(x) \rightarrow y \text{ when } x \rightarrow x^*.$$

Clearly enough, one always has  $\lim \text{int } F(x) \subset \lim \text{ext } F(x)$ ; when these two sets are equal, the common set is the *limit* of  $F(x)$  when  $x \rightarrow x^*$ .

**Remark 5.1** The above concepts are classical but one usually speaks of  $\limsup$  and  $\liminf$ . The reason is that the  $\lim \text{ext}$  [resp.  $\lim \text{int}$ ] is the largest [resp. smallest] cluster set for the order " $\subset$ ". Such a terminology is however misleading, since this order does not generalize " $\leq$ ".

For example, with  $X = \{1, 1/2, \dots, 1/k, \dots\}$  (and  $x^* = 0$ ) what are the  $\limsup$  and  $\liminf$  of the sequence  $\{(-1)^k\}$  for  $k \rightarrow +\infty$ ? With the classical terminology, there are two contradictory answers:

– If  $\{(-1)^k\}$  is considered as a sequence of *numbers* in the ordered set  $(\mathbb{R}, \leq)$ , then

$$\limsup (-1)^k = 1 \quad \text{and} \quad \liminf (-1)^k = -1.$$

– If  $\{(-1)^k\}$  is considered as a sequence of *singletons* in the ordered set  $(2^{\mathbb{R}}, \subset)$ , then

$$\limsup \{(-1)^k\} = \{-1, +1\} \quad \text{and} \quad \liminf \{(-1)^k\} = \emptyset.$$

□

Beware that the above *limes* may cover somewhat pathological behaviours. Take for example the multifunction

$$]0, +\infty[ \ni t \longmapsto F(t) := \{0, 1/t\} \subset \mathbb{R}.$$

Then  $\lim_{t \downarrow 0} \text{ext}_{t \downarrow 0} F(t) = \{0\}$ , a set which does not reflect the intuitive idea that a  $\lim \text{ext}$  should connote. Note: in this example,  $\Delta_H[F(t), \{0\}] = e[F(t)/\{0\}] \rightarrow +\infty$  when  $t \downarrow 0$ . The same pathological behaviour of the Hausdorff distance occurs in the following example:

$$F(t) := [0, 1/t] \quad \text{for } t > 0 \quad \text{and} \quad F(0) = [0, +\infty[.$$

Then  $F(0) = \lim_{t \downarrow 0} F(t)$  but  $\Delta_H[F(t), F(0)] \equiv +\infty$ .

**5.3** The multifunction  $F$  is said to be bounded-valued, closed-valued, convex-valued etc. when the sets  $F(x)$  are bounded, closed, convex etc. In order to avoid the nasty situations mentioned above, a convenient property is *local boundedness*: we say that the multifunction  $F$  is locally bounded near  $x^*$  when:

$$\begin{aligned} \text{For some neighborhood } N \text{ of } x^* \text{ and bounded set } B \subset \mathbb{R}^n, \\ N \subset \text{dom } F \quad \text{and} \quad F(N) \subset B. \end{aligned} \tag{5.1}$$

If  $F$  is locally bounded near every  $x^*$  in a set  $S$ , we say that  $F$  is locally bounded on  $S$ . Then a multifunction  $F$  satisfying (5.1) is

– *outer semi-continuous* at  $x^*$  when

$$\lim_{x \rightarrow x^*} \text{ext } F(x) \subset F(x^*),$$

– *inner semi-continuous* at  $x^*$  when

$$F(x^*) \subset \lim_{x \rightarrow x^*} \text{int } F(x).$$

– *continuous* when it is both outer and inner semi-continuous.

When  $F$  is closed-valued, these definitions can be made more handy thanks to (5.1), namely: for all  $\varepsilon > 0$ , there is a neighborhood  $N(x^*)$  such that  $x \in N(x^*)$  implies

$$\begin{aligned} F(x) \subset F(x^*) + B(0, \varepsilon) &\quad [\text{outer semi-continuity}] \\ F(x^*) \subset F(x) + B(0, \varepsilon). &\quad [\text{inner semi-continuity}] \end{aligned} \tag{5.2}$$

It is straightforward to check that (5.2) has an equivalent in terms of excesses:

$$\begin{aligned} e[F(x)/F(x^*)] &\leq \varepsilon & [\text{outer semi-continuity}] \\ e[F(x^*)/F(x)] &\leq \varepsilon . & [\text{inner semi-continuity}] \end{aligned}$$

In words, outer semi-continuity at  $x^*$  means: all the points in  $F(x)$  are close to  $F(x^*)$  if the varying  $x$  is close to the fixed  $x^*$ . When moving away from  $x^*$ ,  $F$  does not expand suddenly. Inner semi-continuity is the other way round:  $F(x)$  does not explode when the varying  $x$  reaches  $x^*$ . If the mapping is actually single-valued, both definitions coincide with that of a continuous function at  $x^*$ . In practice, outer semi-continuity is frequently encountered, while inner semi-continuity is less natural.

**5.4** Finally, we mention a situation in which limits and continuity have a natural definition: when  $X$  is an ordered set, and  $x \mapsto F(x)$  is *nested*. For example, take  $X = \mathbb{R}^+$  and let  $t \mapsto F(t)$  satisfy

$$t \geq t' > 0 \implies F(t) \subset F(t') .$$

Then the set

$$\lim_{t \downarrow 0} F(t) := \text{cl} \cup \{F(t) : t > 0\}$$

coincides with the limit of  $F$  defined in §5.2.

## 6 A Bird's Eye View of Measure Theory and Integration

We equip  $\mathbb{R}^n$  with the Lebesgue measure, denoted by  $\mu$ , or  $\mu_n$  if necessary. The integrability (with respect to  $\mu$ ) of a function  $f$  is always understood in Lebesgue's sense; the integral of  $f$  is denoted by  $\int f d\mu$ , but we also use the more familiar notation

$$\int_{\mathbb{R}^n} f(x) dx \quad \text{or} \quad \int_{\mathbb{R}^n} f(\xi^1, \dots, \xi^n) d\xi^1 \cdots d\xi^n .$$

For example, if a measurable set  $S \subset \mathbb{R}^n$  is bounded, and if its characteristic function is  $\chi_S$  (1 on  $S$ , 0 elsewhere), we have  $\int_{\mathbb{R}^n} \chi_S(x) dx = \mu(S)$ .

**6.1** For univariate functions, *Lebesgue's differentiation theorem* is as follows. If  $f : [a, b] \rightarrow \mathbb{R}$  is increasing, then  $f$  is differentiable almost everywhere (i.e. the set where  $f$  has no derivative is of zero measure). Recall that a function as above need not be continuous, but has only countably many discontinuities.

A function  $f : [a, b] \rightarrow \mathbb{R}$  is *absolutely continuous* when it satisfies the following property. For any  $\varepsilon > 0$ , there exists  $\delta$  such that, for any countable collection of disjoint subintervals  $[a_k, b_k] \subset [a, b]$  with  $\sum_k (b_k - a_k) \leq \delta$ , one has  $\sum_k |f(b_k) - f(a_k)| \leq \varepsilon$ .

The fundamental property of absolutely continuous functions is that, except possibly on a set of measure zero, they have a (finite) derivative and

$$f(b) - f(a) = \int_a^b f'(t) dt .$$

In words, an absolutely continuous function is the integral of its derivative, which exists at sufficiently many points. Another way of saying the same thing: a function  $f$  is absolutely continuous (on  $[a, b]$ ) if, and only if, it is the indefinite integral of an integrable function; there is  $g$  integrable such that

$$f(y) = f(x) + \int_x^y g(t) dt \quad (a \leq x < y \leq b).$$

**6.2 Fatou's Lemma.** Let  $\{f_k\}$  be a sequence of nonnegative measurable functions on  $\mathbb{R}^n$  and define the function  $x \mapsto f(x) := \liminf_{k \rightarrow +\infty} f_k(x)$ . Then the following inequality holds:

$$\int_{\mathbb{R}^n} f(x) dx \leq \liminf_{k \rightarrow +\infty} \int_{\mathbb{R}^n} f_k(x) dx;$$

“integrating nonnegative functions is a lower semi-continuous operation”.

**Fubini's Theorem** of successive integrations. Let  $f : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}$  be a measurable function and assume that one of the integrals

$$\int_{\mathbb{R}^{p+q}} |f(x, y)| dx dy, \quad \int_{\mathbb{R}^p} \left[ \int_{\mathbb{R}^q} |f(x, y)| dy \right] dx, \quad \int_{\mathbb{R}^q} \left[ \int_{\mathbb{R}^p} |f(x, y)| dx \right] dy$$

is finite (this is therefore an integrability assumption). Then

$$\int_{\mathbb{R}^{p+q}} f(x, y) dx dy = \int_{\mathbb{R}^p} \left[ \int_{\mathbb{R}^q} f(x, y) dy \right] dx = \int_{\mathbb{R}^q} \left[ \int_{\mathbb{R}^p} f(x, y) dx \right] dy.$$

We retain in particular that, for an integrable function, the above formula of successive integrations is true.

A frequent use of this theorem is when  $f$  is the characteristic function of some set  $S \subset \mathbb{R}^p \times \mathbb{R}^q$  (and this is precisely what is needed in this book). Then  $f(x, \cdot)$  is the characteristic function of the “slice” of  $S$  along  $x$ ; this is the set

$$S_x := \{y \in \mathbb{R}^q : (x, y) \in S\}.$$

For example, Fubini's Theorem tells us that, if  $S$  is bounded and measurable, the measure of  $S$  (in  $\mathbb{R}^{p+q}$ ) can be computed by integrating the measure of  $S_x$  (in  $\mathbb{R}^q$ ):

$$\mu_{p+q}(S) = \int_{\mathbb{R}^p} \mu_q(S_x) dx.$$

Thus:

- If  $S_x$  is of zero measure for (almost) all  $x$ , then  $S$  is of zero measure; this is a common way of showing that a set is of zero measure.
- If  $S$  is of zero measure, then *almost all* its slices are of zero measure (draw pictures in  $\mathbb{R}^2$  to be convinced that the word “almost” is essential).

Finally, the integrable character of a function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , as well as the value of its integral, are independent of any system of coordinates in  $\mathbb{R}^n$ .

A good reference for this Section 6 is for example [172, Chap. 13]; we also suggest [181] more particularly for §6.1.

## Bibliographical Comments

Some notion of convexity appeared already in Archimedes' works. Closer to us, modern convexity theory resulted in several branches, using often the same tools or basic concepts, but with problems to solve of diverse nature: geometric convexity is one example. In the present book, it is the *variational* aspect, or the relationship with *continuous optimization* that we have stressed.

The development of convexity during the last half-century owes much to W. Fenchel (1905-1988), J.-J. Moreau (1923-), R.T. Rockafellar (1935-). Fenchel was very “geometrical”; Moreau, according to his own words, did applied mechanics: he “applied Mechanics to Mathematics”; while the concept of “dual problem” is a constant leading thread for Rockafellar. Besides mechanics, one should not forget that convexity comes naturally into play in another branch of science: thermodynamics. There, “convexifying” a function (passing from  $f$  to  $\bar{co} f$ ) is a common operation. The works of the physicist J.W. Gibbs (1839-1903) were a benchmark in this respect: read A.S. Wightman’s introduction to [82]: “Convexity and the notion of equilibrium state in thermodynamics and statistical mechanics”.

For the reader wanting to start a library, here are some suggested books:

- First of all, *the* reference book to keep on the shelf, as far as convex analysis in finite dimension is concerned, is [159].
- Convex sets and functions in infinite dimension: [131], [49, Chaps I–III], [160], [12], [8, Part I], [33], [78], [97, Chaps VI–VII], [143], [27].
- Convexity and mathematical economics: [132], [7], [8].
- Convexity in variational problems: [3], [49], [81], [177], [184].
- Convexity and approximation theory: [77], [97].
- Convexity in statistics, in statistical mechanics: [13], [51].
- Use of convexity in nonsmooth analysis [164], [37], [166].

We now give some comments for more detail on subjects treated in the present volume.

**Chapter I.** Convex functions of one real variable have a fairly old history, which followed that of modern analysis. These functions play a role in fields as different as: functional analysis (construction of Birnbaum-Orlicz spaces), probability theory (when using Young functions in martingale theory), graph theory (optimization of flows in a network). The introduction of the contemporary presentation, with functions assuming the value  $+\infty$  (§1.3), was influenced by [158, §2].

For an illustration of the function  $0 < x \mapsto xf(1/x)$ , with  $f$  convex (Theorem 1.1.6), see [15]. The convergence result announced in Proposition 2.2.3 can be found for example in [172, Chap. 13, §12]. See [58, p. 23] for the mean-value theorem in inequality form, used in the proof of Theorem 5.3.1.

Let  $u : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  be continuous, strictly increasing, satisfying  $u(0) = 0$ , and call  $v$  its inverse function. Then

$$sx \leq \int_0^x u(\alpha) d\alpha + \int_0^s v(\beta) d\beta \quad \text{for all } x \geq 0 \text{ and } s \geq 0.$$

This inequality, due to W.H. Young (1912), is a particular case of Fenchel's inequality: take  $f(x) = \int_0^x u(\alpha) d\alpha$ , so that  $f^*(s) = \int_0^s v(\beta) d\beta$ . Besides, it contains the essential part of it (the “functions”  $\partial f$  and  $\partial f^*$  are increasing, so to speak); hence the name Young-Fenchel for inequality (6.1.1). This inequality will be seen in a multi-dimensional context in the second volume of this book, Chap. X.

For a numerical calculation of the conjugate  $f^*$ , or rather  $(f + I_{[a,b]})^*$ , in one dimension, and via an algorithm connoting the fast Fourier transform, see [28]. Iterated convolutions of probability laws play a central role in probability calculus, for which the Fourier transform is the main tool. Likewise, inf-convolutions of the objective function are fundamental for dynamic programming, and there it is the transformation of Fenchel (conjugacy) that comes into play. This analogy is explained and used in [156].

Finally, we mention a peculiarity of the univariate inf-convolution: even when  $f$  and  $g$  are  $C^\infty$  (or even polynomials),  $f \downarrow g$  is of class  $C^6$  but need not be of class  $C^7$ . This is due to [87], where the following example can be found:

$$f(x) = \frac{1}{4}x^4 \quad \text{and} \quad g(x) = \frac{1}{6}x^6.$$

Then

$$(f \downarrow g)(x) = \frac{1}{6}x^6 - \frac{3}{4}|x|^{20/3} + h(x),$$

with  $h \in C^7(\mathbb{R})$ . Note that  $20/3 < 7$ !

**Chapter II.** Our exposition is quite close in spirit to [46], an excellent textbook with a large part devoted to the important field of least squares. See also [56], which contains a treatment of the constrained case.

The steepest-descent method is traditionally attributed to A. Cauchy [34]. To solve a system of equations, say  $F(x) = 0$ , he designed the gradient method minimizing the squared norm  $F(x)^\top F(x)$ . It is interesting to mention that he was worried by the speed of convergence, and proposed to graft Newton's method at the end of the algorithm. Another interesting remark is that he claimed convergence, based on the following argument: we have a sequence  $\{\delta_k\}$  of positive numbers which is decreasing; hence  $\delta_k \rightarrow 0$  (needless to say, the property was nonetheless true; we mention here that Cauchy is considered as one of the most rigorous mathematicians of his time). Theorem 2.2.4 comes from [144, §6.1].

For Newton's method, a classical designation is “Newton-Raphson-Kantorovich”. Conjugate gradients are due to the fundamental paper [69], which should

still be carefully read by anyone working in numerical analysis. As for quasi-Newton methods, the seminal paper is [41] by W.C. Davidon, a physicist who was not so concerned by Hessian operators, but rather by covariance matrices. His work was not published in an international journal until 1991: see [42] and its lively “belated preface”. However [41] was quickly popularized in the mathematical community by [57]; then, for a decade or two, the vast majority of papers in nonlinear numerical optimization dealt with quasi-Newton methods.

For views of line-searches close to those exposed here, see for example the little-known [186], [105], [128], and also the first edition of [56]. It should be mentioned that the modern tendency goes towards the so-called trust-region technique, surveyed in [127]. In Chap. XV of the present book, a few words will be said on the connection between this technique, line-searches, and nonsmooth optimization. The counter-example alluded to in Remark 2.4.4 is due to [150]; but it uses an exact line-search, and it is not clear whether a descent test resembling (3.2.1) is satisfied by that counter-example. For some more views on safe programming of a numerical algorithm (§3.4) see also [126].

**Chapter III.** The first systematic study of convexity (in finite dimension) is due to H. Minkowski (1864-1909); most ideas on the subject can be found in his works, at least in seminal form. Theorem 1.3.6 of Carathéodory (1873-1950) goes back also to the very beginning of the XX<sup>th</sup> Century. A proof of the Fenchel-Bunt Theorem 1.3.7 can be found in [50, Thm 18(ii)], or [83, Lemma B.2.2]. Along the lines of these results, we mention the following theorem (of Shapley-Folkman). Let  $S_1, \dots, S_k$  be subsets of  $\mathbb{R}^n$  and  $S := S_1 + \dots + S_k$  (we know that  $\text{co } S = \text{co } S_1 + \dots + \text{co } S_k$ ). Any  $x \in \text{co } S$  can be expressed as  $x_1 + \dots + x_k$ , with  $x_i \in \text{co } S_i$ , and the set of  $i$  such that  $x_i \notin S_i$  having at most  $n$  elements. This theorem and Carathéodory’s can be viewed as particular cases of a more general result ([5, Lemma I]), relating the dimension of a face of  $C$  exposed by  $s$  and that of  $A(C)$  ( $A$  being affine) exposed by  $As$ . Minkowski’s Theorem 2.3.4 has a generalization to infinite dimension, due to Krein and Milman: a compact convex set  $C$  in a Hausdorff locally convex vector space is the closed convex hull of its extreme points. This explains that Minkowski’s theorem often appears under the banner “theorem of Krein-Milman”.

Moreau’s Theorem 3.2.5 goes back to 1962: [129]. For developments around the Minkowski-Farkas lemma and the associated historical context, consult for example [151], [169]. The use of separation theorems to obtain multipliers in nonlinear programming, and their development through the ages, are recorded in [149]. The directional differentiability Property 5.3.5 of  $p_C$  at  $x \in C$  and the formula  $p'_C(x, \cdot) = p_{T_C(x)}(\cdot)$  are found in [189, p. 300] or [121, Prop. 2]. Note that  $p_C$  need not have a directional derivative at an  $x \notin C$ . For this, additional properties on  $C$  are required.

We have totally neglected here combinatorial aspects in the study of convex sets, and convex geometry. These subjects are treated in [98], [169], [66]. To know more on closed convex polyhedra, we recommend [29].

**Chapter IV.** The variational formulation of the sum of the  $m$  largest eigenvalues of a symmetric matrix  $A$ , as the support function of  $\Omega = \{Q : Q^\top Q = I_m\}$  (§1.3(e)), is due to Ky Fan (1949). Incidentally, it can be shown that  $\text{co } \Omega$  is nothing more than the set of positive semi-definite matrices  $A$  satisfying  $\text{tr } A = m$  and  $\lambda_1(A) \leq 1$ .

The function  $\varphi_S$  of Example 2.1.4 was introduced and studied towards the end of the sixties by E. Asplund, see for example [6]. The operator  $(A_1^{-1} + A_2^{-1})^{-1}$ , constructed in Example 2.3.8, is called the parallel sum of  $A_1$  and  $A_2$ ; this parallel addition appeared for the first time in [4]. For more on the variational approach of this operation, see [120].

The Lipschitzian extension described in Proposition 3.1.4 uses essentially the Lipschitz property of  $f$  on  $C$ , convexity of  $f$  being present just to ensure the convexity of the extension ([70]). Actually, such a procedure goes back at least to Baire and Hausdorff. It must not be confused with the regularization-approximation technique based on the inf-convolution with the norm, alluded to in Proposition I.2.2.4(j), and which will also appear in our §XI.3.4. This last technique (comparable to the Moreau-Yosida regularization of Proposition I.2.2.4(i), using the squared norm) was pointed out and studied in the convex case in [71, §2].

The proof of Theorem 4.2.3 is that of [157, §44]. This last work contains interesting complements on convex functions. We recall that univariate second-order differentiation of convex functions is studied in our §I.5, where we prove (the univariate version of) Alexandroff's Theorem 4.3.4. For the multi-dimensional case, a proof somewhat technical but pedagogical and readable, can be found in the Appendix of [38].

**Chapter V.** The concept of support function has its roots in the works of Minkowski, who considered the three-dimensional case. However, the note [79] of L. Hörmander, written in a general context of locally convex topological spaces, was extremely influential in modern developments.

Let  $\mathbf{H}$  be the set of positively homogeneous functions from  $\mathbb{R}^n$  to  $\mathbb{R}$ . This is a vector space, which can be equipped with the norm (assumed finite for simplicity)

$$\mathbf{H} \ni h \mapsto \|h\| := \sup_{\|x\|=1} |h(x)|.$$

In  $\mathbf{H}$ , we find the convex cone  $\mathbf{K}$  of (finite-valued) sublinear functions  $\sigma : \mathbb{R}^n \rightarrow \mathbb{R}$ . Besides, consider the set  $\mathbf{C}$  of nonempty compact convex sets in  $\mathbb{R}^n$ . The isomorphism of §3

$$\mathbf{C} \ni C \mapsto \sigma_C \in \mathbf{K} \subset \mathbf{H}$$

is sometimes called Radström's embedding.

Formula (3.3.9), expressing the Hausdorff distance  $\Delta_H$  between two sets  $S$  and  $S'$  with the help of their support functions, appears in [79]. We mention that distance functions can also be used: in fact,

$$\Delta_H(S, S') = \sup_{x \in \mathbb{R}^n} |\mathbf{d}_S(x) - \mathbf{d}_{S'}(x)|.$$

Both formulae are applicable to some extent to unbounded sets  $S$  and  $S'$ .

**Chapter VI.** Various names appear in 1963 to denote a vector  $s$  satisfying (1.2.1): R.T. Rockafellar in his thesis (1963) calls  $s$  “a differential of  $f$  at  $x$ ”; it is J.-J. Moreau who, in a Note aux Comptes-Rendus de l’Académie des Sciences (Paris, 1963), introduces for  $s$  the word “sous-gradient”. Our existence proof in (§1.4) of such a subgradient is due to [26]. This is one of the possible arguments for proving the Hahn-Banach Theorem (see [114] for example), which can also be used in a more general context of Lipschitzian functions, as in [85].

In the years 1965–70, various calculus rules concerning sup-functions (§4.4) started to emerge. The time was ripe and several authors contributed to the subject: B.N. Pschenichnyi, V.L. Levin, R.T. Rockafellar, A. Sotskov, . . . who worked in the field and used various assumptions. However, the most elaborated results are due to M. Valadier. In particular, the idea of considering almost active indices appeared in [178]; the counter-example of (4.4.10) is extracted from [179]. Speaking of counter-examples, K.C. Kiwiel found (4.4.7) and Remark 4.5.4 comes from [86]. Generally speaking, our assumptions in this Section 4.4 (finite-dimensional context, finite-valued functions) are more restrictive than those used by most of the above-mentioned authors; however, they allow more refined statements and less technical proofs.

The work presented in §5.1 on the maximal eigenvalue has a complete extension to the sum  $f_m(M)$  of the  $m$  largest eigenvalues of a symmetric matrix  $M$ . As indicated in our preceding comments on Chap. IV,  $\partial f_m(0)$  is the set of positive semi-definite matrices  $A$  such that  $\text{tr } A = m$  and  $\lambda_1(A) \leq 1$ . A general formula for the subdifferential is then:

$$\partial f_m(M) = \{A \in \partial f_m(0) : \langle\langle A, M \rangle\rangle = f_m(M)\},$$

i.e.  $\partial f_m(M)$  is the face of  $\partial f_m(0)$  exposed by  $M$ . More explicit formulae can be given; see [76], [140], and the references therein.

Theorem 6.3.1 (statement and proof) is a convex adaptation of a more general result ([36] or [37, §2.5]): if  $f$  is locally Lipschitzian,  $\text{co } \gamma f(x)$  is called by F.H. Clarke the generalized gradient of  $f$  at  $x$ , whose support function happens to be

$$d \mapsto f^\circ(x, d) := \limsup \left\{ \frac{f(x' + td) - f(x')}{t} : x' \rightarrow x, t \downarrow 0 \right\}.$$

Note the (crucial) perturbation “ $x' \rightarrow x$ ” in the above difference quotient: it makes  $f^\circ(x, \cdot)$  convex. When  $f$  is itself convex,  $f^\circ$  is just the ordinary  $f'$ .

**Chapter VII.** The work of H.W. Kuhn and A.W. Tucker, published in 1951, as well as W. Karush’s M.Sc. thesis (1939), can be viewed as the historical roots of what are called the KKT conditions. Naturally, these authors worked in a differentiable context, though; and they did not consider the constraint-qualifications condition presented in this chapter. The assumption of M. Slater goes back to the same period (1950).

The Lagrange multipliers, their uses and interpretations, constraint-qualification assumptions such as BCQ, the concept of exact penalty, all these themes are encountered in non-convex analysis (smooth or not). On the other hand, the connection between Lagrange multipliers and saddle-points of the Lagrangian, the (global) sensitivity of a problem with respect to perturbations, the mini-maximization approach

rely crucially on convexity. Mathematical economists contributed a lot to these, and use them eagerly. For an analysis of the historical developments, consult [93], [149], [151], [165], [169].

The link between exact penalty and multipliers, with the non-differentiable view adopted here, is inspired by [19]; see also [25]. Numerical approaches based on the John condition (see (iii) at the end of §3.2) is not so common, we mention [99]. Existence results for saddle-points of convex-concave functions on a product of compact convex sets go back to S. Kakutani (1941) and M. Sion (1957-58).

**Chapter VIII.** The finite minimax problem has its champion: V.F. Demjanov ([44]). He devised in [45, Chap. III, §5] the first known counter-example to convergence, which involved nonlinear functions  $f_j$ , so we preferred to borrow from [187] our piecewise affine function of §2.2.

The steepest-descent method considered in this chapter corresponds to the gradient projection method for constrained optimization problems, coming from [167], [144]. As explained in [104], a second general class of algorithms is also convenient: the so-called sequential quadratic programming approach, due to B.N. Pshenichnyi [155], [154]. Nowadays the near totality of software for nonlinear programming is based on it, in conjunction with quasi-Newton techniques. For an exhaustive study of differential inclusions (§3.4) and their applications in economics, see the monograph [9].

## References

1. Aizerman, M.A., Braverman, E.M., Rozonoer, L.I.: The probability problem of pattern recognition learning and the method of potential functions. *Automation and Remote Control* **25**,9 (1964) 1307–1323.
2. Alexeev, V., Galeev, E., Tikhomirov, V.: *Recueil de Problèmes d'Optimisation*. Mir, Moscow (1984).
3. Alexeev, V., Tikhomirov, V., Fomine, S.: *Commande Optimale*. Mir, Moscou (1982).
4. Anderson Jr., W.N., Duffin, R.J.: Series and parallel addition of matrices. *J. Math. Anal. Appl.* **26** (1969) 576–594.
5. Artstein, Z.: Discrete and continuous bang-bang and facial spaces or: look for the extreme points. *SIAM Review* **22**,2 (1980) 172–185.
6. Asplund, E.: Differentiability of the metric projection in finite-dimensional Euclidean space. *Proc. Amer. Math. Soc.* **38** (1973) 218–219.
7. Aubin, J.-P.: *Optima and Equilibria: An Introduction to Nonlinear Analysis*. Springer, Berlin Heidelberg (1993).
8. Aubin, J.-P.: *Mathematical Methods of Game and Economic Theory*. North-Holland (1982) (revised edition).
9. Aubin, J.-P., Cellina, A.: *Differential Inclusions*. Springer, Berlin Heidelberg (1984).
10. Auslender, A.: *Optimisation, Méthodes Numériques*. Masson, Paris (1976).
11. Auslender, A.: Numerical methods for nondifferentiable convex optimization. In: *Nonlinear Analysis and Optimization*. Math. Prog. Study **30** (1987) 102–126.
12. Barbu, V., Precupanu, T.: *Convexity and Optimization in Banach Spaces*. Sijthoff & Noordhoff (1982).
13. Barndorff-Nielsen, O.: *Information and Exponential Families in Statistical Theory*. Wiley & Sons (1978).
14. Bellman, R.E., Kalaba, R.E., Lockett, J.: *Numerical Inversion of the Laplace Transform*. Elsevier (1966).
15. Ben Tal, A., Ben Israel, A., Teboulle, M.: Certainty equivalents and information measures: duality and extremal principles. *J. Math. Anal. Appl.* **157** (1991) 211–236.
16. Berger, M.: *Geometry I, II (Chapters 11, 12)*. Springer, Berlin Heidelberg (1987).
17. Berger, M.: Convexity. *Amer. Math. Monthly* **97**,8 (1990) 650–678.
18. Berger, M., Gostiaux, B.: *Differential Geometry: Manifolds, Curves and Surfaces*. Springer, New York (1990).
19. Bertsekas, D.P.: Necessary and sufficient conditions for a penalty method to be exact. *Math. Prog.* **9** (1975) 87–99.
20. Bertsekas, D.P.: *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press (1982).

21. Bertsekas, D.P.: Convexification procedures and decomposition methods for nonconvex optimization problems. *J. Optimization Th. Appl.* **29**, 2 (1979) 169–197.
22. Bertsekas, D.P., Mitter, S.K.: A descent numerical method for optimization problems with nondifferentiable cost functionals. *SIAM J. Control* **11**, 4 (1973) 637–652.
23. Best, M.J.: Equivalence of some quadratic programming algorithms. *Math. Prog.* **30**, 1 (1984) 71–87.
24. Bihain, A.: Optimization of upper semi-differentiable functions. *J. Optimization Th. Appl.* **4** (1984) 545–568.
25. Bonnans, J.F: Théorie de la pénalisation exacte. *Modélisation Mathématique et Analyse Numérique* **24**, 2 (1990) 197–210.
26. Borwein, J.M.: A note on the existence of subgradients. *Math. Prog.* **24** (1982) 225–228.
27. Borwein, J.M., Lewis, A.: *Convexity, Optimization and Functional Analysis*. Wiley Interscience – Canad. Math. Soc. (in preparation).
28. Brenier, Y.: Un algorithme rapide pour le calcul de transformées de Legendre-Fenchel discrètes. *Note aux C.R. Acad. Sci. Paris* **308** (1989) 587–589.
29. Brøndsted, A.: *An Introduction to Convex Polytopes*. Springer, New York (1983).
30. Brøndsted, A., Rockafellar, R.T.: On the subdifferentiability of convex functions. *Proc. Amer. Math. Soc.* **16** (1965) 605–611.
31. Brousse, P.: *Optimization in Mechanics: Problems and Methods*. North-Holland (1988).
32. Cansado, E.: Dual programming problems as hemi-games. *Management Sci.* **15**, 9 (1969) 539–549.
33. Castaing, C., Valadier, M.: *Convex Analysis and Measurable Multifunctions*. Lecture Notes in Mathematics, vol. 580. Springer, Berlin Heidelberg (1977).
34. Cauchy, A.: Méthode générale pour la résolution des systèmes d'équations simultanées. *Note aux C. R. Acad. Sci. Paris* **25** (1847) 536–538.
35. Cheney, E.W., Goldstein, A.A.: Newton's method for convex programming and Tchebycheff approximation. *Numer. Math.* **1** (1959) 253–268.
36. Clarke, F.H.: Generalized gradients and applications. *Trans. Amer. Math. Soc.* **205** (1975) 247–262.
37. Clarke, F.H.: *Optimization and Nonsmooth Analysis*. Wiley & Sons (1983), reprinted by SIAM (1990).
38. Crandall, M.G., Ishii, H., Lions, P.-L.: User's guide to viscosity solutions of second order partial differential equations. *Bull. Amer. Math. Soc.* **27**, 1 (1992) 1–67.
39. Crouzeix, J.-P.: A relationship between the second derivative of a convex function and of its conjugate. *Math. Prog.* **13** (1977) 364–365.
40. Dantzig, G.B. Wolfe, P.: A decomposition principle for linear programs. *Oper. Res.* **8** (1960) 101–111.
41. Davidon, W.C.: Variable metric method for minimization. AEC Report ANL5990, Argonne National Laboratory (1959).
42. Davidon, W.C.: Variable metric method for minimization. *SIAM J. Optimization* **1** (1991) 1–17.
43. Dedieu, J.-P.: Une condition nécessaire et suffisante d'optimalité en optimisation non convexe et en calcul des variations. Séminaire d'Analyse Numérique, Univ. Paul Sabatier, Toulouse (1979–80).
44. Demjanov, V.F.: Algorithms for some minimax problems. *J. Comp. Syst. Sci.* **2** (1968) 342–380.
45. Demjanov, V.F., Malozemov, V.N.: *Introduction to Minimax*. Wiley & Sons (1974).

46. Dennis, J., Schnabel, R.: *Numerical Methods for Constrained Optimization and Non-linear Equations*. Prentice Hall (1983).
47. Dubois, J.: Sur la convexité et ses applications. *Ann. Sci. Math. Quebec* **I**,1 (1977) 7–31.
48. Dubuc, S.: *Problèmes d'Optimisation en Calcul des Probabilités*. Les Presses de l'Université de Montréal (1978).
49. Ekeland, I., Temam, R.: *Convex Analysis and Variational Problems*. North-Holland, Amsterdam (1976).
50. Eggleston, H.G.: *Convexity*. Cambridge University Press, London (1958).
51. Ellis, R.S.: *Entropy, Large Deviations and Statistical Mechanics*. Springer, New York (1985).
52. Everett III, H.: Generalized Lagrange multiplier method for solving problems of optimum allocation of resources. *Oper. Res.* **11** (1963) 399–417.
53. Fenchel, W.: Convexity through the ages. In: *Convexity and its Applications* (P.M. Gruber and J.M. Wills, eds.). Birkhäuser, Basel (1983) 120–130.
54. Fenchel, W.: Obituary for the death of —. Det Kongelige Danske Videnskabernes Selskabs Aarbk (Oversigten) [Yearbook of the Royal Danish Academy of Sciences] (1988–89) 163–171.
55. Feuer, A.: An implementable mathematical programming algorithm for admissible fundamental functions. PhD. Thesis, Columbia Univ. (1974).
56. Fletcher, R.: *Practical Methods of Optimization*. Wiley & Sons (1987).
57. Fletcher, R., Powell, M.J.D.: A rapidly convergent method for minimization. *The Computer Journal* **6** (1963) 163–168.
58. Flett, T.M.: *Differential Analysis*. Cambridge University Press (1980).
59. Fukushima, M.: A descent algorithm for nonsmooth convex programming. *Math. Prog.* **30**,2 (1984) 163–175.
60. Geoffrion, A.M.: Duality in nonlinear programming: a simplified application-oriented development. *SIAM Review* **13**,11 (1971) 1–37.
61. Gilbert, J.C., Lemaréchal,C.: Some numerical experiments with variable-storage quasi-Newton algorithms. *Math. Prog.* **45** (1989) 407–435.
62. Goffin, J.-L., Haurie, A., Vial, J.-Ph.: Decomposition and nondifferentiable optimization with the projective algorithm. *Management Sci.* **38**,2 (1992) 284–302.
63. Gorni, G.: Conjugation and second-order properties of convex functions. *J. Math. Anal. Appl.* **158**,2 (1991) 293–315.
64. Griewank, A., Rabier, P.J.: On the smoothness of convex envelopes. *Trans. Amer. Math. Soc.* **322** (1990) 691–709.
65. Grinold, R.C.: Lagrangian subgradients. *Management Sci.* **17**,3 (1970) 185–188.
66. Gritzmann, P., Klee, V.: Mathematical programming and convex geometry. In: *Handbook of Convex Geometry*, Elsevier, North-Holland (1993) 627–674.
67. Gruber, P.M.: History of convexity. In: *Handbook of Convex Geometry*, Elsevier, North-Holland (1993), 1–15.
68. Held, M., Karp, R.M.: The traveling-salesman problem and minimum spanning trees. *Math. Prog.* **1**,1 (1971) 6–25.
69. Hestenes, M.R., Stiefel, M.R.: Methods of conjugate gradients for solving linear systems. *J. Res. NBS* **49** (1959) 409–436.
70. Hiriart-Urruty, J.-B.: Extension of Lipschitz functions. *J. Math. Anal. Appl.* **77** (1980) 539–554.
71. Hiriart-Urruty, J.-B.: Lipschitz  $r$ -continuity of the approximate subdifferential of a convex function. *Math. Scand.* **47** (1980) 123–134.

72. Hiriart-Urruty, J.-B.:  $\varepsilon$ -subdifferential calculus. In: *Convex Analysis and Optimization* (J.-P. Aubin and R. Vinter, eds.). Pitman (1982), pp. 43–92.
73. Hiriart-Urruty, J.-B.: Limiting behaviour of the approximate first order and second order directional derivatives for a convex function. *Nonlinear Anal. Theory, Methods & Appl.* **6**, 12 (1982) 1309–1326.
74. Hiriart-Urruty, J.-B.: When is a point  $x$  satisfying  $\nabla f(x) = 0$  a global minimum of  $f$ ? *Amer. Math. Monthly* **93** (1986) 556–558.
75. Hiriart-Urruty, J.-B.: Conditions nécessaires et suffisantes d’optimalité globale en optimisation de différences de fonctions convexes. *Note aux C.R. Acad. Sci. Paris* **309**, I (1989) 459–462.
76. Hiriart-Urruty, J.-B., Ye, D.: Sensitivity analysis of all eigenvalues of a symmetric matrix. Preprint Univ. Paul Sabatier, Toulouse (1992).
77. Holmes, R.B.: *A Course on Optimization and Best Approximation*. Lecture Notes in Mathematics, vol. 257. Springer, Berlin Heidelberg (1972).
78. Holmes, R.B.: *Geometrical Functional Analysis and its Applications*. Springer, Berlin Heidelberg (1975).
79. Hörmander, L.: Sur la fonction d’appui des ensembles convexes dans un espace localement convexe. *Ark. Mat.* **3**, 12 (1954) 181–186.
80. Ioffe, A.D., Levin, V.L.: Subdifferentials of convex functions. *Trans. Moscow Math. Soc.* **26** (1972) 1–72.
81. Ioffe, A.D., Tikhomirov, V.M.: *Theory of Extremal Problems*. North-Holland (1979).
82. Israel, R.B.: *Convexity in the Theory of Lattice Gases*. Princeton University Press (1979).
83. Karlin, S.: *Mathematical Methods and Theory in Games, Programming and Economics*. Mc Graw-Hill, New York (1960).
84. Kelley, J.E.: The cutting plane method for solving convex programs. *J. SIAM* **8** (1960) 703–712.
85. Kim, K.V., Nesterov, Yu.E., Cherkassky, B.V.: The estimate of complexity of gradient computation. *Soviet Math. Dokl.* **27**, 6 (1984) 1306–1309.
86. Kiselman, C.O.: How smooth is the shadow of a smooth convex body? *J. London Math. Soc.* (2) **33** (1986) 101–109.
87. Kiselman, C.O.: Smoothness of vectors sums of plane convex sets. *Math. Scand.* **60** (1987), 239–252.
88. Kiwiel, K.C.: An aggregate subgradient method for nonsmooth convex minimization. *Math. Prog.* **27** (1983) 320–341.
89. Kiwiel, K.C.: *Methods of Descent for Nondifferentiable Optimization*. Lecture Notes in Mathematics, vol. 1133. Springer, Berlin Heidelberg (1985).
90. Kiwiel, K.C.: A survey of bundle methods for nondifferentiable optimization. In: *Proceedings, XIII. International Symposium on Mathematical Programming*, Tokyo (1988).
91. Kiwiel, K.C.: Proximity control in bundle methods for convex nondifferentiable minimization. *Math. Prog.* **46**, 1 (1990) 105–122.
92. Kiwiel, K.C.: A tilted cutting plane proximal bundle method for convex nondifferentiable optimization. *Oper. Res. Lett.* **10** (1991) 75–81.
93. Kuhn, H.W.: Nonlinear programming: a historical view. *SIAM-AMS Proceedings* **9** (1976) 1–26.
94. Kutateladze, S.S.: Changes of variables in the Young transformation. *Soviet Math. Dokl.* **18**, 2 (1977) 545–548.
95. Kutateladze, S.S.: Convex  $\varepsilon$ -programming. *Soviet Math. Dokl.* **20** (1979) 391–393.
96. Kutateladze, S.S.:  $\varepsilon$ -subdifferentials and  $\varepsilon$ -optimality. *Sib. Math. J.* (1981) 404–411.

97. Laurent, P.-J.: *Approximation et Optimisation*. Hermann, Paris (1972)
98. Lay, S.R.: *Convex Sets and their Applications*. Wiley & Sons (1982).
99. Lebedev, B.Yu.: On the convergence of the method of loaded functional as applied to a convex programming problem. *J. Num. Math. and Math. Phys.* **12** (1977) 765–768.
100. Lemaréchal, C.: An algorithm for minimizing convex functions. In: *Proceedings, IFIP74* (J.L. Rosenfeld, ed.). Stockholm (1974), pp. 552–556.
101. Lemaréchal, C.: An extension of Davidon methods to nondifferentiable problems. In: *Nondifferentiable Optimization* (M.L. Balinski, P. Wolfe, eds.). *Math. Prog. Study* **3** (1975) 95–109.
102. Lemaréchal, C.: Combining Kelley's and conjugate gradient methods. In: *Abstracts, IX. Intern. Symp. on Math. Prog.*, Budapest (1976).
103. Lemaréchal, C.: Nonsmooth optimization and descent methods. *Research Report* **78,4** (1978) IIASA, 2361 Laxenburg, Austria.
104. Lemaréchal, C.: Nonlinear programming and nonsmooth optimization: a unification. *Rapport Laboria* **332** (1978) INRIA.
105. Lemaréchal, C.: A view of line-searches. In: *Optimization and Optimal Control* (A. Auslender, W. Oettli, J. Stoer, eds.). *Lecture Notes in Control and Information Sciences*, vol. 30. Springer, Berlin Heidelberg (1981), pp. 59–78.
106. Lemaréchal, C.: Constructing bundle methods for convex optimization. In: *Fermat Days 85: Mathematics for Optimization* (J.-B. Hiriart-Urruty, ed.). North-Holland Mathematics Studies **129** (1986) 201–240.
107. Lemaréchal, C.: An introduction to the theory of nonsmooth optimization. *Optimization* **17** (1986) 827–858.
108. Lemaréchal, C.: Nondifferentiable optimization. In: *Handbook in OR & MS*, Vol. 1 (G.L. Nemhauser et al., eds.). Elsevier, North-Holland (1989), pp. 529–572.
109. Lemaréchal, C., Mifflin, R.: Global and superlinear convergence of an algorithm for one-dimensional minimization of convex functions. *Math. Prog.* **24,3** (1982) 241–256.
110. Lemaréchal, C., Nemirovskij, A.S., Nesterov, Yu.E.: New variants of bundle methods. *Math. Prog.* **69** (1995) 111–147.
111. Lemaréchal, C., Zowe, J.: Some remarks on the construction of higher order algorithms in convex optimization. *Appl. Math. Optimization* **10** (1983) 51–68.
112. Lion, G.: Un savoir en voie de disparition: la convexité. *Singularité* **2,10** (1991) 5–12.
113. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large-scale optimization. *Math. Prog.* **45** (1989) 503–528.
114. Luenberger, D.G.: *Optimization by Vector Space Methods*. Wiley & Sons (1969).
115. Magnanti, T.L.: Fenchel and Lagrange duality are equivalent. *Math. Prog.* **7** (1974) 253–258.
116. Marcotte, P., Dussault, J.P.: A sequential linear programming algorithm for solving monotone variational inequalities. *SIAM J. Control Opt.* **27** (1989) 1260–1278.
117. Marsten, R.E.: The use of the boxstep method in discrete optimization. In: *Nondifferentiable Optimization* (M.L. Balinski, P. Wolfe, eds.). *Math. Prog. Study* **3** (1975) 127–144.
118. Martí, J.: *Konvexe Analysis*. Birkhäuser, Basel (1977).
119. Martinet, B.: Régularisation d'inéquations variationnelles par approximations successives. *Revue Franc. Rech. Opér.* **R3** (1970) 154–158.
120. Mazure, M.-L.: L'addition parallèle d'opérateurs interprétée comme inf-convolution de formes quadratiques convexes. *Modélisation Math. Anal. Numér.* **20** (1986) 497–515.

121. Mc Cormick, G.P., Tapia, R.A.: The gradient projection method under mild differentiability conditions. *SIAM J. Control* **10**,1 (1972) 93–98.
122. Mifflin, R.: Semi-smooth and semi-convex functions in constrained optimization. *SIAM J. Control Opt.* **15**,6 (1977) 959–972.
123. Mifflin, R.: An algorithm for constrained optimization with semi-smooth functions. *Math. Oper. Res.* **2**,2 (1977) 191–207.
124. Mifflin, R.: A modification and an extension of Lemaréchal's algorithm for nonsmooth minimization. In: *Nondifferential and Variational Techniques in Optimization* (D.C. Sorensen, J.B. Wets, eds.). *Math. Prog. Study* **17** (1982) 77–90.
125. Minoux, M.: *Programmation Mathématique: Théorie et Algorithmes* I, II. Dunod, Paris (1983).
126. Moré, J.J.: Implementation and testing of optimization software. In: *Performance Evaluation of Numerical Software* (L.D. Fosdick, ed.). North-Holland (1979).
127. Moré, J.J.: Recent developments in algorithms and software for trust region methods. In: *Mathematical Programming, the State of the Art* (A. Bachem, M. Grötschel, B. Korte, eds.). Springer, Berlin Heidelberg (1983), pp. 258–287.
128. Moré, J.J., Thuente, D.J.: Line search algorithms with guaranteed sufficient decrease. *ACM Transactions on Math. Software* **20** (1994) 286–307.
129. Moreau, J.-J.: Décomposition orthogonale d'un espace hilbertien selon deux cônes mutuellement polaires. *C.R. Acad. Sci. Paris* **255** (1962) 238–240.
130. Moreau, J.-J.: Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France* **93** (1965) 273–299.
131. Moreau, J.-J.: *Fonctionnelles Convexes*. Lecture notes, Séminaire “Equations aux dérivées partielles”, Collège de France, Paris (1966).
132. Moulin, H., Fogelman-Soulé, F.: *La Convexité dans les Mathématiques de la Décision*. Hermann, Paris (1979).
133. Nemirovskij, A.S., Yudin, D.B.: *Problem Complexity and Method Efficiency in Optimization*. Wiley-Interscience (1983).
134. Nesterov, Yu.E.: Minimization methods for nonsmooth convex and quasiconvex functions. *Matekon* **20** (1984) 519–531.
135. Niven, I.: *Maxima and Minima Without Calculus*. Dolciani Mathematical Expositions **6** (1981).
136. Nurminskii, E.A.: On  $\epsilon$ -subgradient mappings and their applications in nondifferentiable optimization. Working paper **78**,58 (1978) IIASA, 2361 Laxenburg, Austria.
137. Nurminskii, E.A.:  $\epsilon$ -subgradient mapping and the problem of convex optimization. *Cybernetics* **21**,6 (1986) 796–800.
138. Nurminskii, E.A.: Convex optimization problems with constraints. *Cybernetics* **23**,4 (1988) 470–474.
139. Nurminskii, E.A.: A class of convex programming methods. *USSR Comput. Maths Math. Phys.* **26**,4 (1988) 122–128.
140. Overton, M.L., Womersley, R.S.: Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Math. Prog.* **62** (1993) 321–358.
141. Penot, J.-P.: Subhessians, superhessians and conjugation. *Nonlinear Analysis: Theory, Methods and Appl.* **23** (1994) 689–702.
142. Peressini, A.L., Sullivan, F.E., Uhl, J.J.: *The Mathematics of Nonlinear Programming*. Springer, New York (1988).

143. Phelps, R.R.: *Convex Functions, Monotone Operators and Differentiability*. Lecture Notes in Mathematics, vol. 1364. Springer, Berlin Heidelberg (1989, new edition in 1993).
144. Polak, E.: *Computational Methods in Optimization*. Academic Press, New York (1971).
145. Poljak, B.T.: A general method for solving extremum problems. Soviet Math. Dokl. **174**,8 (1966) 33–36.
146. Poljak, B.T.: Minimization of unsmooth functionals. USSR Comput. Maths Math. Phys. **9** (1969) 14–29.
147. Popova, N.K., Tarasov, V.N.: A modification of the cutting-plane method with accelerated convergence. In: *Nondifferentiable Optimization: Motivations and Applications* (V.F. Demjanov, D. Pallaschke, eds.). Lecture Notes in Economics and Mathematical Systems, vol. 255. Springer, Berlin Heidelberg (1984), pp. 284–190.
148. Ponstein, J.: Applying some modern developments to choosing your own Lagrange multipliers. SIAM Review **25**,2 (1983) 183–199.
149. Pourciau, B.H.: Modern multiplier rules. Amer. Math. Monthly **87** (1980), 433–452.
150. Powell, M.J.D.: Nonconvex minimization calculations and the conjugate gradient method. In: *Numerical Analysis* (D.F. Griffiths ed.). Lecture Notes in Mathematics, vol. 1066. Springer, Berlin Heidelberg (1984), pp. 122–141.
151. Prekopa, A.: On the development of optimization theory. Amer. Math. Monthly **87** (1980) 527–542.
152. Pshenichnyi, B.N.: *Necessary Conditions for an Extremum*. Marcel Dekker (1971).
153. Pshenichnyi, B.N.: Nonsmooth optimization and nonlinear programming. In: *Nonsmooth Optimization* (C. Lemaréchal, R. Mifflin, eds.), IIASA Proceedings Series 3, Pergamon Press (1978), pp. 71–78.
154. Pshenichnyi, B.N.: *Methods of Linearization*. Springer, Berlin Heidelberg (1993).
155. Pshenichnyi, B.N., Danilin, Yu.M.: *Numerical Methods for Extremal Problems*. Mir, Moscow (1978).
156. Quadrat, J.-P.: Théorèmes asymptotiques en programmation dynamique. C.R. Acad. Sci. Paris, **311**, Série I (1990) 745–748.
157. Roberts, A.W., Varberg, D.E.: *Convex Functions*. Academic Press (1973).
158. Rockafellar, R.T.: Convex programming and systems of elementary monotonic relations. J. Math. Anal. Appl. **19** (1967) 543–564.
159. Rockafellar, R.T.: *Convex Analysis*. Princeton University Press (1970).
160. Rockafellar, R.T.: *Convex Duality and Optimization*. SIAM regional conference series in applied mathematics (1974).
161. Rockafellar, R.T.: Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Math. Oper. Res. **1**,2 (1976) 97–116.
162. Rockafellar, R.T.: Lagrange multipliers in optimization. SIAM-AMS Proceedings **9** (1976) 145–168.
163. Rockafellar, R.T.: Solving a nonlinear programming problem by way of a dual problem. Symposia Mathematica **XIX** (1976) 135–160.
164. Rockafellar, R.T.: *The Theory of Subgradients ad its Applications to Problems of Optimization: Convex and Nonconvex Functions*. Heldermann, West-Berlin (1981).
165. Rockafellar, R.T.: Lagrange multipliers and optimality. SIAM Review **35** (1993) 183–238.
166. Rockafellar, R.T., Wets, R.J.-B.: *Variational Analysis* (in preparation).
167. Rosen, J.B.: The gradient projection method for nonlinear programming; part I: linear constraints. J. SIAM **8** (1960) 181–217.

168. Schramm, H., Zowe, J.: A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results. *SIAM J. Opt.* **2** (1992) 121–152.
169. Schrijver, A.: *Theory of Linear and Integer Programming*. Wiley-Interscience (1986).
170. Seeger, A.: Second derivatives of a convex function and of its Legendre-Fenchel transformate. *SIAM J. Opt.* **2**,**3** (1992) 405–424.
171. Shor, N.Z.: *Minimization Methods for Nondifferentiable Functions*. Springer, Berlin Heidelberg (1985).
172. Smith, K.T.: *Primer of Modern Analysis*. Springer, New York (1983).
173. Stoer, J., Witzgall, C.: *Convexity and Optimization in Finite Dimension I*. Springer, Berlin Heidelberg (1970).
174. Strang, G.: *Introduction to Applied Mathematics*. Wellesley – Cambridge Press (1986).
175. Strodiot, J.-J., Nguyen, V.H., Heukemes, N.:  $\varepsilon$ -optimal solutions in nondifferentiable convex programming and some related questions. *Math. Prog.* **25** (1983) 307–328.
176. Tikhomirov, V.M.: Stories about maxima and minima. In: *Mathematical World 1*, Amer. Math. Society, Math. Association of America (1990).
177. Troutman, J.L.: *Variational Calculus with Elementary Convexity*. Springer, New York (1983).
178. Valadier, M.: Sous-différentiels d'une borne supérieure et d'une somme continue de fonctions convexes. *Note aux C. R. Acad. Sci. Paris, Série A* **268** (1969) 39–42.
179. Valadier, M.: *Contribution à l'Analyse Convexe*. Thèse de doctorat ès sciences mathématiques, Paris (1970).
180. Valadier, M.: Intégration de convexes fermés notamment d'épi-graphes. Inf-convolution continue. *Revue d'Informatique et de Recherche Opérationnelle* (1970) 47–53.
181. Van Rooij, A.C.M., Schikhof, W.H.: *A Second Course on Real Functions*. Cambridge University Press (1982).
182. Van Tiel, J.: *Convex Analysis. An Introductory Text*. Wiley & Sons (1984).
183. Wets, R. J.-B.: *Grundlagen konvexer Optimierung*. Lecture Notes in Economics and Mathematical Systems, vol. 137. Springer, Berlin Heidelberg (1976).
184. Willem, M.: *Analyse Convexe et Optimisation*, 3rd edn. Editions CIACO Louvain-La-Neuve (1989).
185. Wolfe, P.: Accelerating the cutting plane method for nonlinear programming. *J. SIAM* **9**,**3** (1961) 481–488.
186. Wolfe, P.: Convergence conditions for ascent methods. *SIAM Review* **11** (1968) 226–235.
187. Wolfe, P.: A method of conjugate subgradients for minimizing nondifferentiable functions. In: *Proceedings, XII. Annual Allerton conference on Circuit and System Theory* (P.V. Kokotovic, E.S. Davidson, eds.). Univ. Illinois at Urbana-Champaign (1974), pp. 8–15.
188. P. Wolfe: A method of conjugate subgradients for minimizing nondifferentiable functions. In: *Nondifferentiable Optimization* (M.L. Balinski, P. Wolfe, eds.). *Math. Prog. Study* **3** (1975) 145–173.
189. Zarantonello, E.H.: Projections on convex sets in Hilbert spaces and spectral theory. In: *Contributions to Nonlinear Functional Analysis*. Academic Press (1971), pp. 237–424.
190. Zeidler, E.: *Nonlinear Functional Analysis and its Applications III. Variational Methods and Optimization*. Springer, New York (1985).

# Index

- active (set), 27, 138, 253, 266, 304, 357
- addition, *see* sum
- adjoint, 228, 263, 392
- affine
  - combination, 94
  - function, 19, 37, 201, 234, 309
  - hyperplane, 88, 223, 243, 289, 344
  - manifold, 88, 94, 102, 117, 296
  - mapping, 91, 159
  - minorant, 23, 148, 150, 241
- affine hull, *see* hull
- affinely independent, 95
- Alexandrov's theorem, 192
- angle, 59, 117, 137
- asymptotic
  - cone, 109, 136, 179, 203, 214
  - function, 179, 205
- barycentric, 95
- biconjugate, 38
- Bouligand's cone, *see* tangent cone
- breadth, 209, 239
- Carathéodory, 98, 112, 125, 171
- catastrophe, 384
- closed
  - convex cone, 130, 133, 134
  - convex function, 17, 38, 44
  - function, 149, 164
  - graph (mapping), 282
  - multifunction, 398
  - set, 92, 110, 111, 218
- closure
  - of a cone, 128
  - of a function, 19, 38, 149, 150, 171
  - of a set, 93, 100, 103, 127
- coercive, 15
  - (0-), 180, 338, 339
- (1-), 41, 181
- compact, compactness, 109, 282
  - and convergence of functions, 177
  - and convergence of gradients, 284
  - and extreme points, 111
  - and Hausdorff topology, 230
  - criterion for, 204
  - criterion for convex-, 109, 180
  - of a conical hull, 102
  - of a convex hull, 100
  - of an Argmin, 182, 334
  - of multipliers, 312
- complementarity slackness, 307
- complexity theory, 138
- computer, computing, 51, 52, 74, 78, 85, 98
- concave, 145
- cone, 89, 198, 304
  - conical combination, 101
  - conical hull, *see* hull
- conjugate function, 159
- conjugate gradient, 376
- constrained optimization problem, 166, 366
- constraint, 138, 234, 279
- contingent cone, *see* tangent cone
- continuous
  - (absolutely), 26, 378, 399
  - (uniformly), 80
- convergence, 250
  - (global), 49
  - (of functions), 177, 207
  - (of gradients and subgradients), 284
  - (of sets), 232
  - (speed of), 59, 60, 65, 69, 394
  - (uniform), 12, 177, 208, 284
- convex combination, 6, 95, 102, 111, 146, 361
- convex hull, *see* hull

- convex multiplier, **6**, 89
- convexification, **44**, 171, 347
- corner point, *see* kink
- critical
  - cone, 255, 256
  - direction, 255, 256, 354, 379
  - point, *see* stationary
- decomposition, **121**, 263, 277
- deconvolution, **166**, 178, 230
- degenerate, 256
- derivative, **22**, 133
  - (directional), **22**, 55, 188, **238**, 243, 250
    - of a projection, 141
    - (one-sided), **21**, 238, 378
    - (second), **34**, 192
  - descent, **52**, **71**, 343
    - direction, **54**, 256, **343**
  - difference quotient, **17**, 161
    - of second order, 33
    - of sets, **110**, 135
  - differential, **394**
  - differential inclusion, 277, 378
  - dimension, **88**, **97**, **105**
    - of a convex set, **103**
  - direction subspace, 88
  - distance, **135**, **153**, **159**, **300**, **396**
    - (subdifferential of), 259
    - between functions, 206
    - between sets, **209**, **230**, **232**, **233**, **397**
  - domain, **8**, **23**, **144**
    - of a multifunction, 397
  - dot-product, **390**
  - dual, duality, **88**, **208**, **327**, **392**
    - variable, 339
    - norm, **221**
    - problem, 350
  - edge, **112**
  - eigenvalue, **155**, **233**, **374**
  - electrical circuit, 165
  - ellipsoid, **154**, **155**, **356**
  - epigraph, **2**, **8**, **15**, **145**, **156**
    - (strict), **2**, **10**, **145**, **156**, **163**, **169**
  - epigraphic hull, *see* hull
  - Euclidean norm, 393
  - Euclidean space, **390**
  - Euler relation, 289
  - excess, **396**
  - exposed (face, point), **114**, **140**, **220**, **241**, **250**, 287
  - extended-valued, **8**, **144**, **388**
  - extreme point, **110**, **111**, **175**, **372**
  - face, **112**, **114**
  - face (exposed), *see* exposed face
  - facet, **112**, **115**, **222**
  - Farkas, 234
  - feasible direction, **135**, **369**
  - Fenchel
    - duality theorem, **43**
    - transformation, **38**
  - Fenchel-Bunt's theorem, 99
  - fixed point, 329
  - form, **392**
    - (linear), **22**, **115**, **195**, **215**
    - (quadratic), **66**, **154**, **165**
      - (square root of), 202
  - Fréchet, 189, 250, 251
  - Fubini, 190, 257, 286, 400
  - Gâteaux, 250, 251
  - gauge, **202**, **220**, **224**, **258**
  - Gauss-Seidel, **57**, **59**
  - gradient, **51**, **185**, **193**, **394**
  - gradient method, 58
  - graph, **2**, **184**, **282**, **397**
  - Hahn-Banach, **122**, **219**, **248**
  - half-space, **88**, **91**, **113**, **126**, **127**, **147**, **150**
  - Hausdorff, *see* distance between sets
  - Hessian, **395**
  - hull
    - (affine), **94**, **103**, **105**, **193**, **209**, **212**, **362**
    - (closed conical), **102**, **131**
    - (closed convex), **43**, **100**, **171**, **197**, **211**, **288**
    - (conical), **101**, **119**, **129**, **303**
    - (convex), **70**, **96**, **111**, **115**, **116**, **286**, **357**
      - of a function, **171**, **172**
    - (epigraphic), **156**, **168**
  - hyperplane, *see* affine and also supporting
  - image-function, **167**, **228**
  - indicator function, **18**, **152**
  - inf-convolution, **10**, **12**, **163**, **168**, **175**, **206**, **235**, **326**
    - (exact), **28**, **163**, **164**, **274**

- infimand, 386
- infimum, 385
  - of a function, 386
- Jacobian, 395
- Jensen's inequality, 6, 146
- Karush-Kuhn-Tucker, 306, 314
- kink, 24, 212, 252, 254, 260
- Lagrange, Lagrangian, 278, 318, 360, 367
  - multiplier, 306, 337, 356
- Lebesgue measure, 189, 399
- Legendre, 38
- level-set, 246
- lim ext, limes exterior, 28, 135, 379, 397
- lim int, limes interior, 397
- line-search, 53, 159, 346
- linear programming, 339
- Lipschitz, 12, 16, 173, 181, 239, 378
- locally bounded, 16, 232, 282, 300, 398
- log-convex, 160
- lower-bound function, 156, 163, 203
- majorize, majorization, 386
- manifold, *see* affine manifold
- marginal function, 45, 167, 168, 273
- mean-value theorem, 4, 26, 165
- minimax, maximin, 333, 357, 360
- minimum, minimum point, 46, 148, 182, 379, 382, 387
  - (global), 253, 293, 387
  - (local), 48, 253, 293
- Minkowski, 92, 111
- minorize, minorization, 386
- monotone operator, 118, 185, 280, 378
  - (strongly), 185
- Moreau's decomposition theorem, 121, 133
- Moreau-Yosida, 13, 28
- multifunction, 23, 30, 232, 397
- Newton, quasi-Newton, 63, 69, 319, 347, 376
- nonexpansive, 116, 118, 233
- normal cone, 136, 220, 245, 255, 295, 348
- normalization, norming, 345
- objective function, 9, 47, 167, 386
- orthant, 90, 119
- orthogonal, 119, 121, 133, 215
- penalty, 298
  - (exact), 300, 371
- perspective-function, 160, 179, 201, 230
- perturbation function, 45, 167
- piecewise affine, 32, 153, 260, 380
- pivot, pivoting, 366
- polar
  - cone, 119, 137, 214, 304, 344
  - set, 221, 223
- polyhedral
  - cone, 127
  - function, 153, 172
- polyhedron (closed convex), 127, 138, 234, 355
- positive (semi-)definite, 392
- positively homogeneous, 14, 56, 179, 197, 289, 315, 345
- primal function, 45, 167
- primal problem, 339, 350
- projection, 70, 116, 141, 350, 353
  - (non-orthogonal), 297
- proper convex function, 144
- quadratic estimate, 33, 36, 184, 193
- quadratic function, *see* form (quadratic)
- quadratic programming, 296, 326, 340, 355
- qualification, 307, 314
- quasi-convex, 145, 180
- Rademacher, 190
- rate of convergence, *see* convergence (speed of)
- recession (cone, function), *see* asymptotic regularization, 11, 375
- relative
  - boundary, 103
  - interior, 103, 150, 212
- saddle-point, 328, 334
- saddle-value, 329
- safeguard-reduction property, 73
- secant method, *see* Newton, quasi-Newton
- selection, 26, 258, 397
- semi-continuous
  - (inner), 232, 283, 398
  - (lower), 17, 148, 232, 388
  - (outer), 232, 283, 398
  - (upper), 177, 232, 283, 388
- semi-infinite programming, 279, 373

- separation, 247, 344
- (proper), 124
- set-valued mapping, *see* multifunction
- shadow, 92, 273
- simplex, *see* unit simplex
- slack variable, 346
- Slater, 245, 309, 311, 338
- slice, 92, 230, 400
- slope, 38, 179, 243, 321, 392
  - (increasing), 4
- star-difference, 93, 109, 166, 230
- star-shaped, 87
- stationary point, 49, 253
- steepest descent, 347
  - direction, 55, 313, 345
- stepsize, 53, 61, 62, 98, 343
- strictly convex, 3, 40, 48, 143, 185, 281
- strongly convex, 13, 143, 154, 183, 185, 255, 280
- strongly monotone, *see* monotone
- subadditive, 198
- subderivative, 22
- subdifferential, subgradient, 239, 241, 243, 286, 288
- sublevel-set, 18, 145, 148, 149, 180, 244, 386
- sublinear function, 197, 238, 247
- sum
  - of epigraphs, 164
  - of functions, 43
  - of infima, 387
- of sets, 92
- supplement (orthogonal), 392
- support function, 122, 127, 208, 242, 258
- support, supporting, 37, 248
  - hyperplane, 113, 225, 355
- supremum, 385
- tangency, 133, 184, 196, 241
- tangent cone, 133, 136, 243, 245
- tangent direction, 134, 361
- tangent hyperplane, 184, 196, 246
- trace
  - of a function, 256, 289
  - of a matrix, 155, 391
- transversality condition, 120, 307
- trust-region, 403
- uniform convergence, *see* convergence
- unit ball, 393
- unit simplex, 6, 88
- unit sphere, 393
- value function, 45, 167
- variational, 116, 165
- vertex, *see* exposed point
- Wolfe, 77, 79
- Yosida, *see* Moreau-Yosida
- zigzag, 364

# Grundlehren der mathematischen Wissenschaften

*A Series of Comprehensive Studies in Mathematics*

---

*A Selection*

208. Lacey: The Isometric Theory of Classical Banach Spaces
209. Ringel: Map Color Theorem
210. Gihman/Skorohod: The Theory of Stochastic Processes I
211. Comfort/Negrepontis: The Theory of Ultrafilters
212. Switzer: Algebraic Topology – Homotopy and Homology
215. Schaefer: Banach Lattices and Positive Operators
217. Stenström: Rings of Quotients
218. Gihman/Skorohod: The Theory of Stochastic Processes II
219. Duvaut/Lions: Inequalities in Mechanics and Physics
220. Kirillov: Elements of the Theory of Representations
221. Mumford: Algebraic Geometry I: Complex Projective Varieties
222. Lang: Introduction to Modular Forms
223. Bergh/Löfström: Interpolation Spaces. An Introduction
224. Gilbarg/Trudinger: Elliptic Partial Differential Equations of Second Order
225. Schütte: Proof Theory
226. Karoubi: K-Theory. An Introduction
227. Grauert/Remmert: Theorie der Steinschen Räume
228. Segal/Kunze: Integrals and Operators
229. Hasse: Number Theory
230. Klingenberg: Lectures on Closed Geodesics
231. Lang: Elliptic Curves. Diophantine Analysis
232. Gihman/Skorohod: The Theory of Stochastic Processes III
233. Stroock/Varadhan: Multidimensional Diffusion Processes
234. Aigner: Combinatorial Theory
235. Dynkin/Yushkevich: Controlled Markov Processes
236. Grauert/Remmert: Theory of Stein Spaces
237. Köthe: Topological Vector Spaces II
238. Graham/McGehee: Essays in Commutative Harmonic Analysis
239. Elliott: Probabilistic Number Theory I
240. Elliott: Probabilistic Number Theory II
241. Rudin: Function Theory in the Unit Ball of  $C^n$
242. Huppert/Blackburn: Finite Groups II
243. Huppert/Blackburn: Finite Groups III
244. Kubert/Lang: Modular Units
245. Cornfeld/Fomin/Sinai: Ergodic Theory
246. Naimark/Stern: Theory of Group Representations
247. Suzuki: Group Theory I
248. Suzuki: Group Theory II
249. Chung: Lectures from Markov Processes to Brownian Motion
250. Arnold: Geometrical Methods in the Theory of Ordinary Differential Equations
251. Chow/Hale: Methods of Bifurcation Theory
252. Aubin: Nonlinear Analysis on Manifolds. Monge-Ampère Equations
253. Dwork: Lectures on  $p$ -adic Differential Equations
254. Freitag: Siegelsche Modulfunktionen
255. Lang: Complex Multiplication
256. Hörmander: The Analysis of Linear Partial Differential Operators I
257. Hörmander: The Analysis of Linear Partial Differential Operators II
258. Smoller: Shock Waves and Reaction-Diffusion Equations
259. Duren: Univalent Functions
260. Freidlin/Wentzell: Random Perturbations of Dynamical Systems

261. Bosch/Güntzer/Remmert: Non Archimedian Analysis – A System Approach to Rigid Analytic Geometry
262. Doob: Classical Potential Theory and Its Probabilistic Counterpart
263. Krasnosel'skii/Zabreiko: Geometrical Methods of Nonlinear Analysis
264. Aubin/Cellina: Differential Inclusions
265. Grauert/Remmert: Coherent Analytic Sheaves
266. de Rham: Differentiable Manifolds
267. Arbarello/Cornalba/Griffiths/Harris: Geometry of Algebraic Curves, Vol. I
268. Arbarello/Cornalba/Griffiths/Harris: Geometry of Algebraic Curves, Vol. II
269. Schapira: Microdifferential Systems in the Complex Domain
270. Scharlau: Quadratic and Hermitian Forms
271. Ellis: Entropy, Large Deviations, and Statistical Mechanics
272. Elliott: Arithmetic Functions and Integer Products
273. Nikol'skii: Treatise on the Shift Operator
274. Hörmander: The Analysis of Linear Partial Differential Operators III
275. Hörmander: The Analysis of Linear Partial Differential Operators IV
276. Liggett: Interacting Particle Systems
277. Fulton/Lang: Riemann-Roch Algebra
278. Barr/Wells: Toposes, Triples and Theories
279. Bishop/Bridges: Constructive Analysis
280. Neukirch: Class Field Theory
281. Chandrasekharan: Elliptic Functions
282. Lelong/Gruman: Entire Functions of Several Complex Variables
283. Kodaira: Complex Manifolds and Deformation of Complex Structures
284. Finn: Equilibrium Capillary Surfaces
285. Burago/Zalgaller: Geometric Inequalities
286. Andrianov: Quadratic Forms and Hecke Operators
287. Maskit: Kleinian Groups
288. Jacod/Shiryev: Limit Theorems for Stochastic Processes
289. Manin: Gauge Field Theory and Complex Geometry
290. Conway/Sloane: Sphere Packings, Lattices and Groups
291. Hahn/O'Meara: The Classical Groups and K-Theory
292. Kashiwara/Schapira: Sheaves on Manifolds
293. Revuz/Yor: Continuous Martingales and Brownian Motion
294. Knus: Quadratic and Hermitian Forms over Rings
295. Dierkes/Hildebrandt/Küster/Wohlrab: Minimal Surfaces I
296. Dierkes/Hildebrandt/Küster/Wohlrab: Minimal Surfaces II
297. Pastur/Figotin: Spectra of Random and Almost-Periodic Operators
298. Berline/Getzler/Vergne: Heat Kernels and Dirac Operators
299. Pommerenke: Boundary Behaviour of Conformal Maps
300. Orlik/Terao: Arrangements of Hyperplanes
301. Loday: Cyclic Homology
302. Lange/Birkenhake: Complex Abelian Varieties
303. DeVore/Lorentz: Constructive Approximation
304. Lorentz/v. Goliček/Makovoz: Constructive Approximation. Advanced Problems
305. Hiriart-Urruty/Lemaréchal: Convex Analysis and Minimization Algorithms I. Fundamentals
306. Hiriart-Urruty/Lemaréchal: Convex Analysis and Minimization Algorithms II. Advanced Theory and Bundle Methods
307. Schwarz: Quantum Field Theory and Topology
308. Schwarz: Topology for Physicists
309. Adem/Milgram: Cohomology of Finite Groups
310. Giacquinta/Hildebrandt: Calculus of Variations I: The Lagrangian Formalism
311. Giacquinta/Hildebrandt: Calculus of Variations II: The Hamiltonian Formalism
312. Chung/Zhao: From Brownian Motion to Schrödinger's Equation
313. Malliavin: Stochastic Analysis
314. Adams/Hedberg: Function Spaces and Potential Theory
315. Bürgisser/Clausen/Shokrollahi: Algebraic Complexity Theory