

Springer Undergraduate Texts
in Mathematics and Technology

SUMAT

Francisco J. Aragón
Miguel A. Goberna
Marco A. López
Margarita M. L. Rodríguez

Nonlinear Optimization



Springer

Springer Undergraduate Texts in Mathematics and Technology

Series editors

H. Holden, Norwegian University of Science and Technology,
Trondheim, Norway
Keri A. Kornelson, University of Oklahoma, Norman, OK, USA

Editorial Board

Lisa Goldberg, University of California, Berkeley, CA, USA
Armin Iske, University of Hamburg, Germany
Palle E. T. Jorgensen, The University of Iowa, Iowa City, IA, USA

Springer Undergraduate Texts in Mathematics and Technology (SUMAT) publishes textbooks aimed primarily at the undergraduate. Each text is designed principally for students who are considering careers either in the mathematical sciences or in technology-based areas such as engineering, finance, information technology and computer science, bioscience and medicine, optimization or industry. Texts aim to be accessible introductions to a wide range of core mathematical disciplines and their practical, real-world applications; and are fashioned both for course use and for independent study.

More information about this series at <http://www.springer.com/series/7438>

Francisco J. Aragón · Miguel A. Goberna ·
Marco A. López · Margarita M. L. Rodríguez

Nonlinear Optimization



Springer

Francisco J. Aragón
Department of Mathematics
University of Alicante
Alicante, Spain

Marco A. López
Department of Mathematics
University of Alicante
Alicante, Spain

Miguel A. Goberna
Department of Mathematics
University of Alicante
Alicante, Spain

Margarita M. L. Rodríguez
Department of Mathematics
University of Alicante
Alicante, Spain

ISSN 1867-5506 ISSN 1867-5514 (electronic)
Springer Undergraduate Texts in Mathematics and Technology
ISBN 978-3-030-11183-0 ISBN 978-3-030-11184-7 (eBook)
<https://doi.org/10.1007/978-3-030-11184-7>

Library of Congress Control Number: 2018966836

Mathematics Subject Classification (2010): 65kxx, 97uxx, 47N10

© Springer Nature Switzerland AG 2019, corrected publication 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book is aimed at upper-level undergraduate students of mathematics and statistics, and graduate students of industrial engineering. We assume that the readers are familiar with real analysis and linear algebra (subjects that are briefly revisited in Chapter 1), and we decline to use sophisticated tools from convex analysis (such as the Rockafellar subdifferential or the Fenchel–Moreau conjugate function) and nonsmooth analysis (such as the Clarke and Mordukhovich subdifferentials). It is not easy to select, among the vast amount of knowledge on the subject, the suitable contents for a textbook on nonlinear optimization. We have included in this book some basic theoretical results (in particular, optimality and duality theorems for different problems) and a representative selection of the many available numerical methods for problems with or without constraints, some of them not using derivatives. Both theoretical results and numerical methods are illustrated with applications to a diverse range of fields.

The contents are organized in two parts that can be used as textbooks for undergraduate courses on convex and nonconvex optimization, respectively. In fact, this book is based on the contents of two undergraduate courses taught at the University of Alicante. These two courses are preceded by a course on linear programming, which is based on the book [42]. Nevertheless, this knowledge is not assumed in the present book. With the exception of Chapter 1, which is preliminary for both parts, Part I and Part II are essentially independent. The sections and subsections marked with an asterisk (*) are mathematically harder than the rest and can be skipped when either the students have a limited background or when the available time is insufficient.

Part I, composed of Chapters 2–4, focuses on the analytic calculation of global minima (required in engineering, operations research, statistics, and economy) and local minima (required in natural sciences, where the equilibrium situations correspond to local minima—or even critical points—of certain functions). Chapters 1 and 2 contain the basic ingredients for the calculus of local minima (optimality conditions for differentiable functions) and global minima (coercivity and convexity) for unconstrained optimization problems. Chapter 3 provides closed formulas for unconstrained optimization problems arising in different fields, while

Chapter 4 deals with unconstrained and constrained convex optimization problems for which local and global minima coincide. For the sake of simplicity, the sections devoted to constrained continuous optimization preclude linear equations and make use of a unique constraint qualification (the Slater condition).

Part II is focused on the numerical calculation of local optima in problems whose solutions cannot be analytically obtained, and it consists of two chapters. Chapter 5 deals with the standard algorithms for unconstrained optimization problems, such as the steepest descent method, Newton's method and variants (trust regions, Gauss–Newton, Levenberg–Marquardt), and other gradient-based methods such as those using conjugate directions (conjugate gradient and quasi-Newton). The local or global convergence of such methods is discussed in detail, as well as the respective rates of convergence (linear, superlinear, and quadratic). This chapter also includes an introduction to methods that do not use derivatives. Finally, Chapter 6 presents, in the first part, an introduction to the so-called penalty and barrier methods. These procedures address the numerical solution of constrained optimization problems by applying the algorithms described in Chapter 5, and they are based on the idea of converting a constrained problem in a sequence of unconstrained ones. The second part of this final chapter is devoted to the optimality conditions for constrained optimization problems, with equality and/or inequality restrictions. The relevance of the so-called constraint qualifications is emphasized as they yield necessary optimality conditions, which can be used as stopping rules for the algorithms. Finally, we show how the optimality conditions give rise to the sequential quadratic programming methods, when Newton's method is applied for solving the associated system of equations.

Definitions, proofs, and numerical methods are illustrated with figures. Moreover, all chapters contain collections of exercises, and the solutions of those exercises that our students usually find harder are provided in a separate appendix at the end of the book.

This is a book on nonlinear optimization for undergraduate students, so it is not intended to provide an exhaustive overview of the available optimality theory and optimization methods. Readers wishing to get a deeper knowledge of the topics included in this book are kindly invited to consult the list of references provided at the end, especially the books [5, 8, 11, 28, 42, 52, 54, 67, 70, 75], whose treatment of different subjects has inspired the approach adopted in some sections. Readers interested in the history of optimization, or in its key role in operations research, may find sources of information in the books [39, 41, 47, 62].

The top five reasons to use this textbook are:

- It only assumes basic knowledge of differential and matrix calculus. All the main concepts, algorithms, and proofs are illustrated with highly explanatory and appealing figures.
- It pays special attention to model building and validation of real problems, and emphasizes the practical advantages of obtaining good reformulations of optimization problems. It shows important applications of optimization models to natural and social sciences, engineering, and data science.

- It provides rigorous optimality conditions. The existence and uniqueness of optimal solutions is analyzed via coercivity and convexity. Dual problems are introduced in order to get lower bounds, sensitivity information (“what if” questions), and stopping rules for primal-dual algorithms.
- It provides an accurate description of the main numerical approaches to solve nonlinear optimization problems. These numerical methods have been chosen with the aim of covering an ample range of techniques. Intentionally, the algorithms described are not implemented in a specific software, as experience shows that this type of choice is highly ephemeral and mainly depends on personal preferences. Nonetheless, five assignments for laboratory sessions of two hours have been included in Section 5.8.
- It has been thoroughly tested. Its preliminary versions have been used for many years as class notes for various undergraduate courses on optimization theory and methods. The exercises have been carefully selected to push the students to a deeper understanding of the main topics in each chapter. Further, a detailed solution to those exercises that our students consider harder is provided, to allow independent study.

We would like to thank our families for their patience and constant support. We are grateful to the anonymous referees and to our editors at Springer, Elizabeth Loew and Razia Amzad, whose suggestions helped us improve this book. We are indebted to Bernardo Cascales, who proposed us to write the initial version of this book in Spanish, and to Jonathan Borwein, who encouraged us to translate the manuscript into English. Our gratitude also goes to all the students of the Mathematics Degree at the University of Alicante who gave us critical remarks on the preliminary versions of this book. Finally, we are thankful to the funding agencies MINECO (Spain) and ERDF (EU) for their financial support along the years. In particular, this book was written as part of the outreach activities of the grant MTM2014-59179-C2-1-P and the research contract RYC-2013-13327 of the first author.

Alicante, Spain
November 2018

Francisco J. Aragón
Miguel A. Goberna
Marco A. López
Margarita M. L. Rodríguez

The original version of this book was inadvertently published with incorrect reference numbers in the frontmatter. The corrections to this book can be found at https://doi.org/10.1007/978-3-030-11184-7_7

Contents

1 Preliminaries	1
1.1 Optimization Models	1
1.1.1 A Brief History of Optimization	1
1.1.2 Building Up Optimization Models	4
1.1.3 Planning Blending Operations	4
1.1.4 Optimization Problems	7
1.1.5 Reformulating Optimization Problems	10
1.2 Basic Calculus Tools	17
1.2.1 Complements of Differential Calculus	17
1.2.2 Complements of Matrix Calculus	27
1.3 Optimality Conditions at Interior Points	33
1.4 Coercivity	38
1.4.1 Coercive Functions	39
1.4.2 The Fundamental Theorem of Algebra*	45
1.5 Exercises	48
Part I Analytical Optimization	
2 Convexity	55
2.1 Convex Sets	55
2.2 Convex Functions of One Variable	62
2.3 Convex Functions of Several Variables	72
2.4 Special Convex Functions	82
2.4.1 Strictly Convex Functions	82
2.4.2 Strongly Convex Functions	83
2.5 Exercises	86
3 Unconstrained Optimization	91
3.1 Unconstrained Quadratic Optimization	92
3.2 Least Squares Regression	94

3.2.1	Linear Regression	95
3.2.2	Polynomial Regression	98
3.3	Least Squares Solutions to Overdetermined Linear Systems	101
3.4	Optimizing Without Derivatives	103
3.4.1	Jensen's Inequalities	103
3.4.2	The Geometric-Arithmetic Inequality	107
3.4.3	Duality in Unconstrained Geometric Optimization [★]	109
3.5	Exercises	113
4	Convex Optimization	117
4.1	The Optimal Set	118
4.2	Unconstrained Convex Optimization	122
4.2.1	The Fermat–Steiner Problem	122
4.2.2	The Fixed-Point Method of Weiszfeld	130
4.2.3	An Application to Statistical Inference	131
4.3	Linearly Constrained Convex Optimization	134
4.3.1	Optimality Conditions	134
4.3.2	Quadratic Optimization	141
4.3.3	Some Closed Formulas	150
4.4	Arbitrarily Constrained Convex Optimization [★]	153
4.4.1	Sensitivity Analysis	153
4.4.2	Optimality Conditions	161
4.4.3	Lagrange Duality	168
4.4.4	Wolfe Duality	170
4.5	A Glimpse on Conic Optimization [★]	172
4.6	Exercises	176

Part II Numerical Optimization

5	Unconstrained Optimization Algorithms	183
5.1	Line Search Methods	183
5.1.1	The Family of Gradient Methods	186
5.1.2	The Stepsize	186
5.2	Convergence of the Line Search Methods	193
5.2.1	Rate of Convergence	197
5.2.2	Convergence Rate of Gradient Methods for Quadratic Forms	201
5.2.3	Gradient Algorithms	207
5.3	Newton's Method	209
5.3.1	Trust Region Methods	212
5.3.2	Least Squares Problems	214
5.4	Newton's Method for Solving Equations	216

5.5	Conjugate Direction Methods	218
5.5.1	Conjugate Gradient Method	222
5.5.2	Quasi-Newton Methods	226
5.6	Derivative-Free Optimization Methods	232
5.6.1	Coordinate Descent Algorithms	233
5.6.2	Nelder and Mead Method	234
5.6.3	Directional Direct Search Methods Using Positive Spanning Sets	236
5.7	Exercises	241
5.8	Computer Exercises	246
5.8.1	Assignment 1	246
5.8.2	Assignment 2	248
5.8.3	Assignment 3	249
5.8.4	Assignment 4	250
5.8.5	Assignment 5	251
6	Constrained Optimization	253
6.1	Penalty and Barrier Methods	253
6.1.1	Penalty Methods	254
6.1.2	Methods Using Exterior Penalty Functions	256
6.1.3	Barrier Methods	260
6.1.4	A Logarithmic Barrier Approach to Linear Programming	264
6.2	Problems with Equality Constraints	266
6.3	Problems with Inequality Constraints	276
6.3.1	Karush–Kuhn–Tucker Optimality Conditions	276
6.3.2	Other Constraint Qualifications [★]	281
6.3.3	Fritz John Optimality Conditions [★]	288
6.4	Problems with Equality and Inequality Constraints	289
6.4.1	Second-Order Optimality Conditions [★]	292
6.5	Sequential Quadratic Programming Methods[★]	297
6.6	Concluding Remarks[★]	302
6.7	Exercises	305
Correction to: Nonlinear Optimization		C1
Solutions to Selected Exercises		311
References		343
Index		347

Nomenclature

$]a, b[$	Open interval between the real numbers a and b
$[a, b]$	Closed interval between the real numbers a and b
\mathbb{R}_+	Set of nonnegative real numbers
\mathbb{R}_{++}	Set of positive real numbers
$\overline{\mathbb{R}}$	Extended real line, page 154
$\min X$	Minimum of $X \subset \mathbb{R}$
$\max X$	Maximum of $X \subset \mathbb{R}$
$\inf X$	Infimum of $X \subset \mathbb{R}$
$\sup X$	Supremum of $X \subset \mathbb{R}$
I_n	$n \times n$ identity matrix
A_k	$k \times k$ submatrix of A which results of taking the first k rows and columns
Δ_k	k th director principal minor of A
A^T	Transpose matrix of A
A^{-1}	Inverse matrix of A
$\det A$	Determinant of the matrix A
$\text{diag}(\lambda_1, \dots, \lambda_n)$	Diagonal matrix with diagonal elements $\lambda_1, \dots, \lambda_n$
$\rho(A)$	Spectral radius of the matrix A , page 31
$\text{cond}(A)$	Condition number of the matrix A , page 31
\mathcal{S}_n	Family of symmetric $n \times n$ real matrices
\mathcal{S}_q^+	Cone of positive semidefinite symmetric matrices
$0_{n \times n}$	$n \times n$ null matrix
0_n	Zero vector of \mathbb{R}^n
e_i	i th vector of the standard basis
1_n	Vector of all ones in \mathbb{R}^n
$\ x\ _1$	Manhattan (or ℓ_1) norm of x , page 15
$\ x\ _\infty$	Chebyshev (or ℓ_∞) norm of x , page 15
$\ x\ $	Euclidean (or ℓ_2) norm of x , page 15
\mathbb{B}	Euclidean closed unit ball
$x^T y$	Scalar product of the vectors x and y

$\{x_k\}$	Sequence in \mathbb{R}^n
\limsup	Upper limit of a sequence
$ X $	Cardinality of X
$\text{int } X$	Topological interior of X
$\text{cl } X$	Closure of X
$\text{bd } X$	Boundary of X
$\text{span } X$	Linear hull of X
$\text{conv } X$	Convex hull of X , page 55
$\text{cone } X$	Conical (convex) hull of X , page 58
K°	(Negative) polar cone of the cone K , page 135
K_p^m	Second-order cone, page 172
$f'(x)$	Derivative of f at x
$\frac{\partial f(x)}{\partial x_i}$	Partial derivative of f with respect to x_i at x
$\frac{\partial^2 f(x)}{\partial x_j \partial x_i}$	Second-order partial derivative of f with respect to x_i and to x_j at x
$f'(x; u)$	Directional derivative of f at x in the direction of u
$f'_+(a)$	Right derivative of f at a , page 62
$f'_-(a)$	Left derivative of f at a , page 62
$\nabla f(x)$	Gradient of f at x , page 19
$\nabla^2 f(x)$	Gradient matrix of f at x , page 26
$J_f(x)$	Hessian matrix of f at x , page 24
$\text{dom } f$	Domain of the function f , pages 17 and 154
$\text{gph } f$	Graph of the function f , page 17
$\text{epi } f$	Epigraph of the function f , page 72
$S_\lambda(f)$	Sublevel set λ of f , page 38
$\text{argmin } f$	Set of points in which f reaches its minimum value
$o(g(x))$	Landau notation, page 24
v_+	Positive part of the vector v , page 120
p_+	Positive part function, page 121
\mathcal{F}	Feasible set multifunction, page 154
ϑ	Value function, page 154
$\mathcal{C}(X)$	Linear space of continuous real-valued functions on X
$\mathcal{C}^m(X)$	Linear space of real functions on X with continuous partial derivatives of order m
F	Feasible set of the problem P , page 8
F^*	Optimal set of the problem P , page 8
$v(P)$	Optimal value of the problem P , page 8
$I(\bar{x})$	Set of active indices at \bar{x} , page 134
$A(\bar{x})$	Active cone at \bar{x} , page 135
$D(\bar{x})$	Feasible direction cone at \bar{x} , page 134

Chapter 1

Preliminaries



This chapter starts with some brief historical overview of optimization followed by four subsections devoted to the formulation and reformulation of optimization problems. The subsequent two sections revisit basic tools of mathematical analysis and matrix analysis, which will be frequently used along the book. Next we revisit the local optimality conditions for functions of several variables (also called multivariate) at interior points of their domains that are usually studied in courses on differential calculus. Then we introduce the concept of coercive function, which is used to prove the existence of global minimum of continuous functions on unbounded domains. Finally, we provide an elementary proof of the fundamental theorem of algebra, based on the coercivity of the modulus of complex polynomials.

1.1 Optimization Models

The method of applied mathematics is based on the construction, analysis, resolution and validation of models. In this section, we consider optimization problems for decision making.

1.1.1 A Brief History of Optimization

It is not clear when the term “optimization” was introduced in mathematics. In the seventeenth century, mathematicians used to speak about calculating maxima and minima (e.g., René Descartes headed in 1637 his famous letter to Marin Mersenne, defending his book *La Dioptrique*, with an ironic “Monsieur votre conseiller *de maximis et minimis*” in allusion to Pierre de Fermat). In 1744, Leonard Euler asserted that “For since the fabric of the universe is most perfect, and is the work of a most wise Creator, nothing whatsoever takes place in the universe in which some relation of maximum and minimum does not appear. Wherefore there is absolutely no doubt

that every effect in the universe can be explained as satisfactorily from final causes, by the aid of the method of maxima and minima, as it can from the effective causes themselves.” (quotation from [71, pp. 76–77]). In vernacular language, the words “maximise”, “minimise”, etc. all first originate in England¹ in the work of Jeremy Bentham, the Victorian utilitarian philosopher, who defined as the fundamental axiom of his philosophy the principle that “it is the greatest happiness of the greatest number that is the measure of right and wrong”. Regarding the popular use of the verb “optimize”, it was first used in England in the sense of “behaving in an optimistic way”, and it was already popular in France by the middle of the nineteenth century by keeping its initial meaning.² In either case, mathematicians have been optimizing well before the introduction of the verb “optimize” in the ordinary language. In fact, the long history of optimization traces back to 3000 years ago and has three key scenarios where the geometric, the analytic, and the numerical paradigms arose.

A. North Africa, around 900 B.C. According to Book IV of Virgil’s Aeneid, princess Dido escaped Tyre when her brother, the king Pygmalion, murdered Dido’s husband in hopes of gaining his wealth. Eventually Dido and her followers arrived to the coast of North Africa where Dido asked the Berber king Iarbas for a small bit of land for a temporary refuge until she could continue her journey, only as much land as could be encompassed by an oxhide. Dido selected a lot (the location of the future Carthage) with form of half-circle whose diameter was a sandy beach while the arc was made with thin strips crafted from the oxhide. History or legend, Dido’s isoperimetric problem seems to be the first documented optimization problem from many posed and solved by the ancient mathematicians (Apollonius, Euclides, Archimedes, etc.) by ad hoc methods of constructive geometry.

B. France, 1638. In 1636 Pierre de Carcavi, a colleague of Pierre de Fermat at the Toulouse Parliament, went to Paris as a royal librarian and made contact with Mersenne, who was playing a key role as a clearinghouse for correspondence between eminent philosophers and scientists at that time. Informed by Carcavi on Fermat’s discoveries on falling bodies, Mersenne contacted Fermat giving rise to a stream of letters on mathematics and its physical applications which include the description of Fermat’s method for determining maxima, minima and tangents to curved lines. Invited by Mersenne to give an opinion on *La Dioptrique*, Fermat concluded that Descartes had not correctly deduced his law of refraction: as light passes the border between two media, the ratio between the angles of incidence and refraction (both angles measured with respect to the line perpendicular to the boundary) equals the ratio of the light speed in the corresponding media. Fermat claimed that the light travels the path that takes the least time and applied his method of maxima, minima and tangents to get the correct formulation of the refraction law by replacing the incidence and refraction angles by the corresponding sinuses. In modern terms, Fermat conceived the differential calculus to solve, in an analytic way (i.e., solely based on pencil and paper) optimization problems involving simple functions. The behavior of

¹J.M. Borwein, personal communication, January 2016.

²Private communication of the note *Optimisation versus Optimisme*, sent by J.-B. Hiriart-Urruty to the group SMAI-MODE in February 2014.

the light illustrates Maupertius' least action metaphysical principle stating that in all natural phenomena a quantity called 'action' tends to be minimized; in other words, nature always acts in the most economic way. This principle, stated in 1746, although criticized by philosophers as Voltaire (at *Le candide*) for its vagueness and implicit deism was finally accepted by the scientific community as, step by step, the mentioned 'actions' were identified for a variety of natural phenomena: potential energy, entropy, free energy, etc. At the same time, mathematicians proposed extensions of Fermat's principle (i.e., necessary optimality conditions) allowing one to solve analytically complex optimization problems arising in experimental sciences: equality constrained optimization problems (Lagrange, in the celestial mechanics setting), optimization of functionals defined on sets of curves (Euler, relative to curves of minimal length on surfaces, called geodesics), etc.

C. USA, 1947. According to George Dantzig's memories [30], on a certain day of 1939 he arrived late to a doctoral lecture given by the famous statistician Neyman at the University of California in Berkeley, so that he copied from the blackboard a list of open problems thinking that they were homework. Dantzig reformulated one of them as a linear optimization problem whose feasible solutions formed a kind of polyhedral set with infinitely many extreme points and edges. To solve such a problem Dantzig conceived the way to jump from an extreme point to an adjacent one by means of algebraic operations. Dantzig considered that this theoretical finding did not deserve publication and gave up temporary his PhD thesis to join the Pentagon. After World War II, Dantzig started to work on planning the postwar complex logistic Pentagon activities, posing linear optimization problems whose feasible set were polyhedral convex sets with large (but finite) sets of extreme points, one of them being the aimed optimal solution. The unique missing ingredient to solve such big problems in practice was a tool able to perform arithmetic operations very fast, which already existed. In fact, the first electronic programmable computer, called ENIAC, had been recently constructed following instructions from John von Neumann, the most famous Western mathematician at that time. Von Neumann supported without hesitation Dantzig's ideas at the first conference on "mathematical programming", a name suggested by the economist Dorfman for the branch of optimization that deals with the numerical methods for those problems that cannot be solved by geometric or analytic methods. Thus, the popular word "program" jumped from economy (where it means "production plan") to mathematics (where it means "optimization problem") and, in turn, from mathematics to computational sciences (where it means "collection of instructions that performs a specific task when executed by a computer").

From an applied perspective, optimization theory is still employed to derive natural laws in physics, chemistry, biology and geology, and closed formulas (optimal solutions expressed in terms of the data) in economic theory and statistics. On the other hand, optimization numerical methods constitute the black box in decision making, with important applications in operations research (also called management science), engineering, data mining, etc.; László Lovász emphasized the importance of linear optimization when, in 1980, stated: "If one would take statistics about which mathematical problem is using up most of the computer time in the world, then [...] the answer would probably be linear programming." The simplex method was indeed selected as one of the ten most crucial algorithms in the twentieth century [23]. From

the research perspective, optimization is at present a very active field of mathematics, as shown by the main database specialized in mathematics: *MathSciNet* (the digital version of the classic journal *Mathematical Reviews*,³ which reviews papers submitted to peer evaluations and books since 1940). In fact, more than 85,000 papers and 6,900 books reviewed in *MathSciNet* are related with optimization, with more than 5,000 papers and 150 books published on 2017.

1.1.2 Building Up Optimization Models

Mathematical models (abstract models that use mathematical language to describe the behavior of systems) arising in natural and social sciences are usually analyzed (e.g., with respect to the existence and uniqueness of solutions) and solved analytically, taking the data as parameters, with the objective of interpreting the obtained solution and its properties.

Mathematical models are also used to make decisions in operations research or engineering on a rational basis, i.e., to map the likely consequences of decisions and choosing the best course of action to take among a set of available alternatives. In order to do this, the decision maker must construct an ad hoc model to be analyzed and solved numerically.

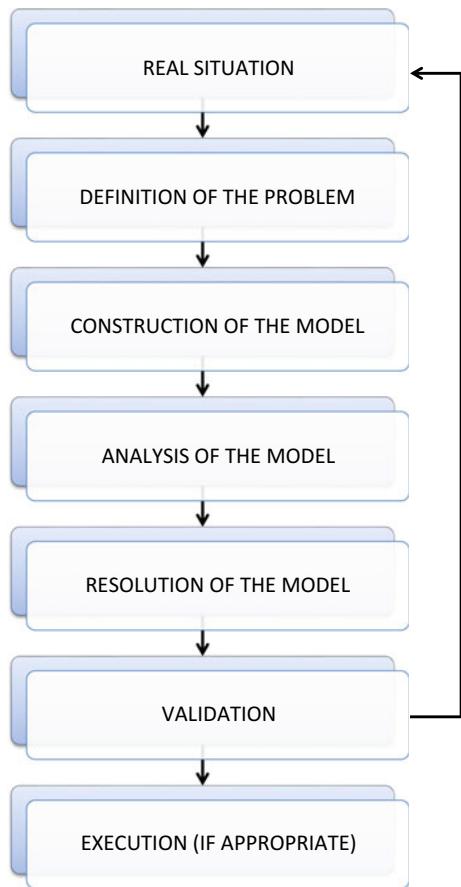
Model building is the key of the method of applied mathematics, which is summarized in Fig. 1.1. When the purpose is making a decision, the first step consists of identifying the objective (or the objectives, if there is more than one) and the restrictions. The resulting optimization problem must be analyzed (Are there feasible solutions? Are there optimal solutions in the sense of Definition 1.1 below? And, if so, how many?) and solved by means of some known method or by a new one. The obtained solution must be validated with respect to the model (checking its feasibility and comparing the objective function on this solution and other feasible solutions obtained via simulation) and with respect to the decision problem. When the solution is not validated, the model must be reformulated as many times as needed until a validated solution is obtained. The following simple example illustrates the method.

1.1.3 Planning Blending Operations

A small oil company produces two types of gasolines, X and Y, from three different oil fields, A, B and C. The transport from the oil fields to the refinery is made as described in the graph of Fig. 1.2. The decision variables x_1, \dots, x_9 represent the amount of oil sent from node to node through the corresponding arc. For instance, the oil extracted at fields A and B share the pipeline in its way to the refinery. The left-hand side arcs also exhibit the unitary extraction costs and the percentage of sulfur (S) for each type of crude oil while the right-hand side arcs show the selling

³<http://www.ams.org/mr-database>.

Fig. 1.1 The method of applied mathematics



price and the legal limit for the percentage of sulfur for each gasoline. We assume that no chemical reaction takes place at the nodes (i.e., that masses are conserved), and it has been decided to produce at most 100 units of X and 200 units of Y.

1.1.3.1 Constructing the Model

- The decision space is \mathbb{R}^9 , whose elements are interpreted as column vectors of components x_1, \dots, x_9 .
- The objective is to maximize the cash flow: $9x_8 + 15x_9 - 6x_1 - 16x_2 - 10x_3$.
- Physical constraints: $x_i \geq 0, i = 1, \dots, 9$.
- Demand constraints: $x_8 \leq 100$ and $x_9 \leq 200$.
- Mass balances at the nodes:

$$\begin{aligned} x_1 + x_2 &= x_4 + x_5, & x_3 &= x_6 + x_7, \\ x_4 + x_6 &= x_8, & x_5 + x_7 &= x_9. \end{aligned}$$

- % of S for the blend obtained from A and B: $\frac{3x_1 + x_2}{x_1 + x_2}$.

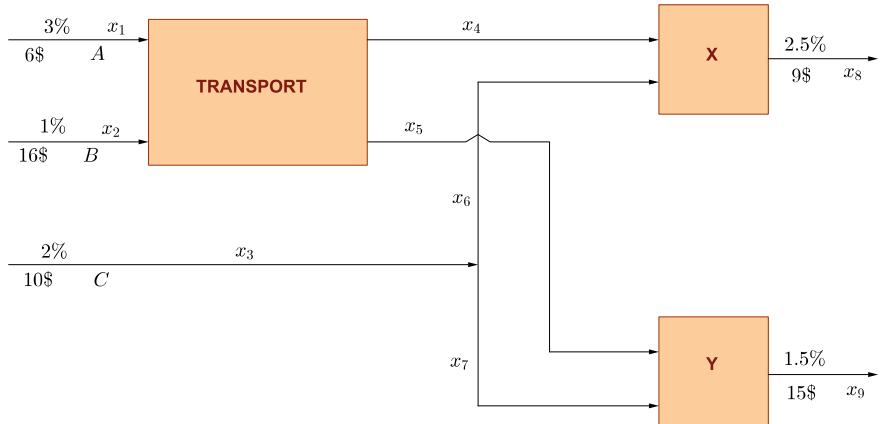


Fig. 1.2 Transport diagram

- Upper bound for the percentage of S in X:

$$\frac{\left(\frac{3x_1 + x_2}{x_1 + x_2}\right)x_4 + 2x_6}{x_4 + x_6} \leq 2.5.$$

- Upper bound for the percentage of S in Y:

$$\frac{\left(\frac{3x_1 + x_2}{x_1 + x_2}\right)x_5 + 2x_7}{x_5 + x_7} \leq 1.5.$$

1.1.3.2 An Optimization Model

After changing the sign of the objective function in order to formulate the problem as a minimization task (as it is customary in optimization) and expressing the last two constraints in terms of quadratic functions, we get the following model:

$$\begin{aligned}
 P : \text{Min } f(x) &= 6x_1 + 16x_2 + 10x_3 - 9x_8 - 15x_9 \\
 \text{s.t. } &x_1x_6 + 3x_2x_4 + x_2x_6 - x_1x_4 \geq 0, \\
 &x_2x_5 - 3x_1x_5 - x_1x_7 - x_2x_7 \geq 0, \\
 &x_1 + x_2 - x_4 - x_5 = 0, \\
 &x_3 - x_6 - x_7 = 0, \\
 &x_4 + x_6 - x_8 = 0, \\
 &x_5 + x_7 - x_9 = 0, \\
 &100 \geq x_8 \geq 0; 200 \geq x_9 \geq 0 \\
 &x_i \geq 0, i = 1, \dots, 7.
 \end{aligned}$$

1.1.3.3 Solving the Optimization Problem

Numerical optimization solvers applied to P (e.g., those available at the free Internet-based NEOS or the commercial software MATLAB) provide feasible solutions satisfying necessary optimality conditions (the KKT conditions in this case) up to some predetermined tolerance. The output depends on the chosen seed (initial point for the corresponding iterative algorithm). Three possible local minima in the sense of Definition 1.3 below can be obtained in this way:

- $\bar{x} = (50, 0, 50, 50, 0, 50, 0, 100, 0)^T$, with value = -100 .
- $\hat{x} = (0, 100, 100, 0, 100, 0, 100, 0, 200)^T$, with value = -400 .
- $\tilde{x} = (0, 0, 200, 0, 0, 0, 200, 0, 200)^T$, with value = -1000 .

1.1.3.4 Validation

It is easy to verify that \bar{x} , \hat{x} and \tilde{x} are feasible solutions of P . Reformulating P by elimination of 4 variables (one per equation at the constraint system), the feasible set of the new problem has dimension $9 - 4 = 5$. One can conclude empirically that \tilde{x} is an optimal solution of P by generating random points in the box $[0, 200]^5$, rejecting the infeasible ones and checking that $f(x) > -1000$ for any feasible solution x generated in this way. However, \tilde{x} is not an admissible decision for the blending problem, as it consists of extracting 200 units of crude oil from field C to produce the same amount of gasoline of type Y with a 2% of sulfur, above the legal limit of 1.5%. Thus, P must be suitably reformulated. How to do it? The answer is left to the reader.

1.1.4 Optimization Problems

Continuous optimization deals with problems of the form

$$\begin{aligned} P : \text{Min } f(x) \\ \text{s.t. } h_i(x) = 0, i = 1, \dots, m, \\ g_j(x) \leq 0, j = 1, \dots, p, \\ x \in C, \end{aligned} \tag{1.1}$$

where “Min” is an abbreviation of the task “Minimize”, $C \subset \mathbb{R}^n$ is a nonempty set called *constraint set*, $f : C \rightarrow \mathbb{R}$ is the objective function, and $h_i, g_j : C \rightarrow \mathbb{R}$, $i = 1, \dots, m$, $j = 1, \dots, p$, are the constraint functions. By default, C is the intersection of the domains of the involved functions (frequently, it is the whole *decision space* \mathbb{R}^n). The problem P is said to be *unconstrained* whenever its unique constraint is $x \in C$ (e.g., $x \in \mathbb{R}^n$).

Definition 1.1 The *feasible set* of the problem P in (1.1) is

$$F := \{x \in C : h_i(x) = 0, i = 1, \dots, m; g_j(x) \leq 0, j = 1, \dots, p\}.$$

The *optimal value* of P is

$$v(P) = \begin{cases} +\infty, & \text{if } F = \emptyset, \\ \inf f(F), & \text{if } f(F) \neq \emptyset \text{ is bounded from below,} \\ -\infty, & \text{if } f(F) \text{ has no lower bound,} \end{cases}$$

in which case P is said to be *inconsistent*, *bounded*, and *unbounded*, respectively. A feasible solution $\bar{x} \in F$ is a *global minimum* of P when $f(\bar{x}) = v(P)$ or, equivalently,

$$f(\bar{x}) \leq f(x), \quad \forall x \in F. \quad (1.2)$$

The problem P is called *solvable* whenever it has global minima. A global minimum $\bar{x} \in F$ is unique when the inequality in (1.2) holds strictly for every feasible solution $x \neq \bar{x}$.

We represent by $F^* = \operatorname{argmin}\{f(x) : x \in F\}$ the *optimal set* of P , i.e., $F^* = \{x \in F : f(x) = v(P)\}$.

We denote by $\mathcal{C}(C)$ the linear space of the real-valued functions that are continuous on C .

Proposition 1.2 (Closedness of the feasible and optimal sets) *If $\emptyset \neq C \subset \mathbb{R}^n$ is closed and all functions involved in problem P in (1.1) are continuous on C , i.e.,*

$$\{f; h_i, i = 1, \dots, m; g_j, j = 1, \dots, p\} \subset \mathcal{C}(C),$$

then F and F^ are closed sets.*

Proof We can assume without loss of generality that F and F^* are nonempty. We also assume that P contains equality and inequality constraints (otherwise the proof can be simplified). We just have to show that both sets contain the limits of the convergent sequences contained in them.

Let $\{x_k\} \subset F$ be such that $\lim_{k \rightarrow \infty} x_k = \bar{x}$ (in short, $x_k \rightarrow \bar{x}$). Since $\{x_k\} \subset C$ and this set is closed, $\bar{x} \in C$. Moreover, for $i = 1, \dots, m$, $h_i(x_k) = 0$; since $x_k \rightarrow \bar{x}$ and $h_i \in \mathcal{C}(C)$, $h_i(x_k) \rightarrow h_i(\bar{x}) = 0$ (the intuitive meaning of the continuity of h_i at \bar{x} is that limit and h_i commute). Analogously, $g_j(\bar{x}) \leq 0$, $j = 1, \dots, p$. Thus, $\bar{x} \in F$.

We now assume that $\{x_k\} \subset F^*$ and $x_k \rightarrow \bar{x}$. On the one hand, since F is closed and $\{x_k\} \subset F$, $\bar{x} \in F$. On the other hand, since $f(x_k) = v(P)$ for all $k \in \mathbb{N}$ and $f \in \mathcal{C}(C)$, we have $f(\bar{x}) = v(P)$, i.e., $\bar{x} \in F^*$. \square

Definition 1.3 A feasible solution $\bar{x} \in F$ is a *local minimum* of P if there exists a neighborhood V of \bar{x} such that

$$f(\bar{x}) \leq f(x), \quad \forall x \in F \cap V, \quad (1.3)$$

and it is a *strict local minimum* of P when the inequality in (1.3) is strict for $x \in F \cap V$ such that $x \neq \bar{x}$.

Analytical optimization deals with methods providing global minima, or at least local minima, of P with pen and paper, using vector, convex, and matrix calculus. In particular cases, it is possible to get “closed formulae” expressing the optimal set F^* in terms of the data which are used in statistics (maximum likelihood estimators), data analysis (regression, principal component analysis, factor analysis), physics, etc. Analytical optimization studies the existence and uniqueness of optimal solutions, i.e., to check whether $F^* \neq \emptyset$ and, if so, to check whether $|F^*| = 1$ ($|F^*|$ represents the *cardinality* of F^* , i.e., the number of global minima).

Numerical optimization is also called *mathematical programming*. Some classes of functions from \mathbb{R}^n to \mathbb{R} that frequently arise in mathematical programming are the following:

- *Affine*: $f(x) = c^T x + b$, with $c \in \mathbb{R}^n$, $b \in \mathbb{R}$ and $c^T x$ representing the scalar product in \mathbb{R}^n ; in particular, f is *linear* when $b = 0$.
- *Quadratic*: $f(x) = \frac{1}{2}x^T Qx - c^T x + b$, with Q being an $n \times n$ symmetric matrix, $c \in \mathbb{R}^n$ and $b \in \mathbb{R}$.
- *Posynomial*: $f(x)$ is a finite sum of monomials of the form $\alpha(x_1^{a_1})(x_2^{a_2}) \dots (x_n^{a_n})$, with $a_1, a_2, \dots, a_n \in \mathbb{R}$ and $\alpha > 0$.

Definition 1.4 Let P be a problem as in (1.1). If $C = \mathbb{R}^n$ and all constraint functions are affine, then P is a *linear optimization* problem when f is linear, and it is a *quadratic optimization* problem when f is quadratic. When $C \neq \mathbb{R}^n$ or at least one of the involved functions is not affine, P is said to be a *nonlinear optimization* problem.

Some nonlinear optimization problems can be solved via linear optimization through a suitable reformulation.

Definition 1.5 An optimization problem P is a *geometric optimization* problem when it can be expressed as

$$\begin{aligned} P : \text{Min } f(x) \\ \text{s.t. } g_j(x) \geq 1, j = 1, \dots, p, \\ x \in \mathbb{R}_{++}^n, \end{aligned}$$

where f, g_1, \dots, g_p are posynomials and \mathbb{R}_{++} represents the set of positive real numbers.

For problems of similar sizes (concept which we do not need to precise here), the diagram in Fig. 1.3 allows to compare the relative difficulty of several classes of optimization problems (the bracketed numbers grow with the difficulty of the corresponding class of problems). Other favorable features for the numerical treatment of nonlinear optimization problems are the differentiability and convexity of

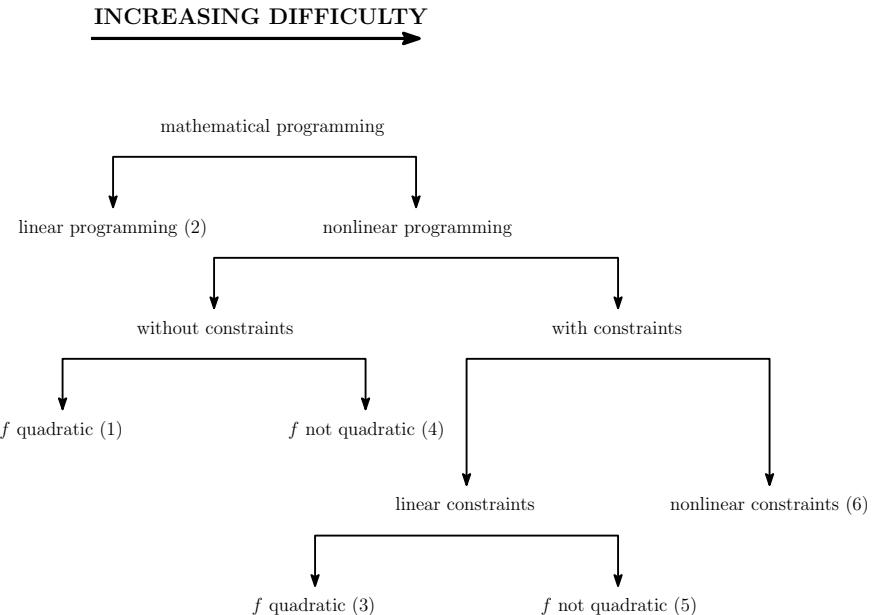


Fig. 1.3 Comparison of the relative difficulty of various classes of optimization problems

the involved functions (in terms of their graphs, the existence of a tangent hyperplane at each point and the property that any arc is below its corresponding chord, respectively). The notion of differentiability will be revised in Subsection 1.2.1, while convexity is the main topic of Chapter 2.

1.1.5 Reformulating Optimization Problems

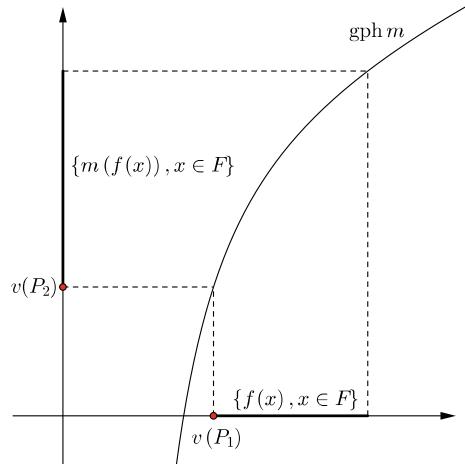
1.1.5.1 Monotone Transformations of Objectives and Constraints

Let P be a problem as in (1.1). A function $m : f(F) \rightarrow \mathbb{R}$ is *increasing* when $y_1 < y_2$ implies $m(y_1) < m(y_2)$ for any pair $y_1, y_2 \in f(F) \subset \mathbb{R}$. If a function m is increasing on $f(F) \subset \mathbb{R}$, then it is injective and its inverse mapping m^{-1} is also increasing on $(m \circ f)(F) \subset \mathbb{R}$.

If $m : f(F) \rightarrow \mathbb{R}$ is increasing (e.g., $m(x) = e^x$, \sqrt{x} , x^3 , $\log x$, $\ln x$, ...) and we denote by P_2 the result of replacing $f(x)$ by $m(f(x))$ in a given optimization problem P_1 , then P_1 and P_2 are equivalent in the sense that they have the same local and global minima and, moreover, $v(P_2) = m(v(P_1))$, see Fig. 1.4.

Analogously, if $m(x)$ is increasing on $g_j(C)$, replacing the constraint $g_j(x) \leq 0$ in P_1 by $m(g_j(x)) \leq m(0)$ provides an equivalent optimization problem P_2 , with $v(P_2) = v(P_1)$.

Fig. 1.4 Reformulating the objective function



Finally, if $m(x)$ es injective on $h_i(C)$ and we replace the constraint $h_i(x) = 0$ in P_1 by $m(h_i(x)) = m(0)$, we also get an equivalent problem.

Example 1.6 We must design a one liter box minimizing the total amount of required cardboard. The decision variables are the dimensions, x_1 , x_2 and x_3 , of the box expressed in decimeters:

$$\begin{aligned} P_1 : \text{Min } & f_1(x) = x_1x_2 + x_1x_3 + x_2x_3 \\ \text{s.t. } & x_1x_2x_3 = 1, \\ & x \in \mathbb{R}_{++}^3. \end{aligned}$$

Observe that the constraint $x_1x_2x_3 = 1$ can be replaced by $x_1x_2x_3 \geq 1$ (as we are then adding dominated feasible solutions), converting P_1 into a geometric optimization problem. Eliminating one decision variable in P_1 , for instance $x_3 = (x_1x_2)^{-1}$, one gets the following equivalent unconstrained geometric optimization problem:

$$\begin{aligned} P_2 : \text{Min } & f(x) = x_1x_2 + \frac{1}{x_1} + \frac{1}{x_2} \\ \text{s.t. } & x \in \mathbb{R}_{++}^2. \end{aligned}$$

Making the change of variables $x_j = e^{y_j}$, $j = 1, 2$, one obtains an equivalent unconstrained nonlinear optimization:

$$P_3 : \text{Min}_{y \in \mathbb{R}^2} f(y) = e^{y_1+y_2} + e^{-y_1} + e^{-y_2}$$

In the next sections we show, using optimality conditions and the coercivity of f (i.e., the fact that $f(x)$ tends to $+\infty$ as x moves away from the origin), that $(0, 0)^T$ is the unique optimal solution of P_3 , so $(1, 1)^T$ is the unique optimal solution of P_2 and $(1, 1, 1)^T$ is the unique optimal solution of P_1 . In subsequent chapters, we shall solve P_1 in different ways.

1.1.5.2 Continuous Formulation of Integer and Mixed Optimization Problems

A 0-1 variable (also called binary or Boolean) becomes continuous by the obvious equivalence

$$x \in \{0, 1\} \Leftrightarrow x(1-x) = 0.$$

An integer variable x taking bounded values on the feasible set can also be replaced by continuous variables as follows. Let $x \in [-M, M]$ on F , with $x \in \mathbb{Z}$. We can write

$$x = y - z, \quad 0 \leq y \leq M, \quad 0 \leq z \leq M, \quad y, z \in \mathbb{Z}.$$

The numbers $0, 1, \dots, 2^{p+1} - 1$ can be expressed as

$$y_0 + y_1 2 + \dots + y_p 2^p, \quad y_i \in \{0, 1\}, \quad i = 0, 1, 2, \dots, p. \quad (1.4)$$

Thus, we can write y as in (1.4) when $M \leq 2^{p+1} - 1$, i.e., $p \geq \log_2(M+1) - 1$. Hence, we can replace x by the 0-1 variables y_0, \dots, y_p and z_0, \dots, z_p through

$$x = (y_0 - z_0) + (y_1 - z_1)2 + \dots + (y_p - z_p)2^p,$$

by just adding the constraints

$$\begin{cases} y_i(1 - y_i) = 0, & i = 0, \dots, p \\ z_i(1 - z_i) = 0, & i = 0, \dots, p \end{cases}.$$

1.1.5.3 Reformulating the Constraints

One can formulate an optimization problem only with equations or only with inequalities. In fact, any equation can be replaced by two inequalities as

$$h(x) = 0 \Leftrightarrow [h(x) \geq 0 \text{ and } h(x) \leq 0],$$

while, analogously, any inequality can be replaced by an equation involving one extra variable as

$$g(x) \leq 0 \Leftrightarrow g(x) + v^2 = 0,$$

where v is a new variable ranging on \mathbb{R} .

Some design problems require certain decision vector to range on sphere or Euclidean closed balls. These problems can be simplified by using trigonometric transformations. We consider, for simplicity, the sphere and the ball centered at the origin O_3 with radius R . Replacing the decision vector (x, y, z) in P by (θ, φ) for the sphere and (θ, φ, r) for the ball by means of the known formulas

$$\begin{cases} x = R \sin \theta \cos \varphi \\ y = R \sin \theta \sin \varphi \\ z = R \cos \theta \end{cases} \quad \text{and} \quad \begin{cases} x = r \sin \theta \cos \varphi \\ y = r \sin \theta \sin \varphi \\ z = r \cos \theta \end{cases},$$

with

$$x^2 + y^2 + z^2 - R^2 = 0 \Leftrightarrow 0 \leq \theta \leq \pi \text{ and } 0 \leq \varphi \leq 2\pi,$$

and

$$x^2 + y^2 + z^2 - R^2 \leq 0 \Leftrightarrow 0 \leq \theta \leq \pi, 0 \leq \varphi \leq 2\pi \text{ and } 0 \leq r \leq R,$$

for the sphere and the ball, respectively, one gets a new optimization problem with a quadratic constraint less and (4 or 6) additional box constraints that can be easily handled.

1.1.5.4 Elimination of Absolute Values in the Objective Function

Differentiability is a desirable property of the objective function. When the objective function is nondifferentiable due to the presence of some addend of the form $\alpha|c^T x + d|$, with $\alpha > 0$, the problem admits an equivalent reformulation with a differentiable objective function.

To make it simple, suppose that our problem P is unconstrained and has only two variables, x and y , and that the objective function f has the form $f(x) = \varphi(x, y) + |x|$, where φ is a differentiable function; that is,

$$P : \text{Min } \varphi(x, y) + |x|.$$

Let us replace x by a difference of two nonnegative variables:

$$x = u - v, \quad u, v \geq 0.$$

Since $u, v \geq 0$, we have $|u - v| \leq u + v$, and equality holds if either $u = 0$ or $v = 0$. Consider now the (constrained) problem

$$\begin{aligned} P_1 : \text{Min } & \varphi(u - v, y) + u + v \\ \text{s.t. } & u, v \geq 0. \end{aligned}$$

Observe that any optimal solution (\bar{u}, \bar{v}) of P_1 must have either \bar{u} or \bar{v} equal to zero. Indeed, let $\delta := \min\{\bar{u}, \bar{v}\}$. If $\delta > 0$, then $(\bar{u} - \delta, \bar{v} - \delta)$ is also feasible, while

$$\begin{aligned} \varphi((\bar{u} - \delta) - (\bar{v} - \delta), y) + \bar{u} - \delta + \bar{v} - \delta &= \varphi(\bar{u} - \bar{v}, y) + \bar{u} + \bar{v} - 2\delta \\ &< \varphi(\bar{u} - \bar{v}, y) + \bar{u} + \bar{v}, \end{aligned}$$

and we obtain a contradiction with the assumed optimality.

Therefore, the optimal solutions of P and P_1 are the same, in the sense that if \bar{x} is optimal for P , then, defining

$$\bar{u} := \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases} \quad \text{and} \quad \bar{v} = \begin{cases} 0, & x \geq 0 \\ -x, & x < 0, \end{cases}$$

we have that (\bar{u}, \bar{v}) is optimal for P_1 , and if (\tilde{u}, \tilde{v}) is optimal for P_1 , then $\tilde{x} = \tilde{u} - \tilde{v}$ is optimal for P . Moreover, the optimal values of both problems must be the same, as any optimal solution (\bar{u}, \bar{v}) of P_1 must have either \bar{u} or \bar{v} equal to zero, which implies $\bar{u} + \bar{v} = |\bar{u} - \bar{v}| = |\bar{x}|$.

Likewise, if $f(x, y) = \varphi(x, y) + \alpha|c^T x + d|$, with $\alpha > 0$, we can introduce two new variables u and v such that

$$c^T x + d = u - v, \quad u, v \geq 0,$$

and replace $\alpha|c^T x + d|$ by $\alpha(u + v)$ in the objective function.

More generally, when the objective function of problem P in (1.1) has the form $f(x) = \varphi(x) + \sum_{k=1}^q \alpha_k |a_k^T x - b_k|$, with φ differentiable and $\alpha_k > 0$ for all k , we can reformulate P by introducing $2q$ new variables u_k, v_k , with corresponding linear constraints $u_k, v_k \geq 0$ and $u_k - v_k = a_k^T x - b_k$, $k = 1, \dots, q$, and replacing the objective function f by $f(x, u, v) = \varphi(x) + \sum_{k=1}^q \alpha_k (u_k + v_k)$.

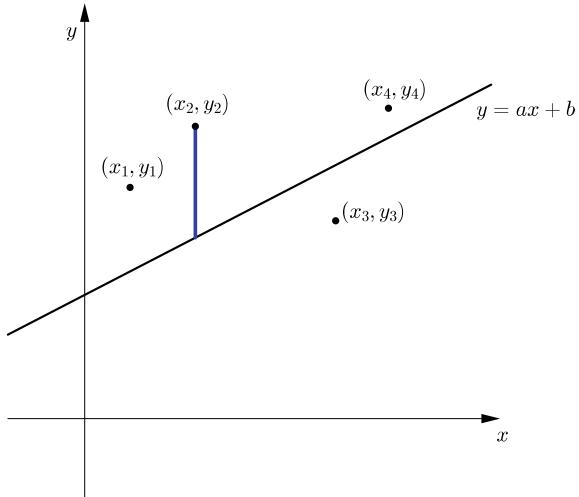
1.1.5.5 ℓ_∞ Linear Regression via Linear Optimization

In statistics, linear regression is an approach for modeling the relationship between a scalar dependent variable y and one or more explanatory variables (or independent variables). We only consider here the case of one explanatory variable x , which is called simple linear regression. Linear regression is the predominant empirical tool in economics. For example, it is widely used in economy to predict consumption spending, inventory investment, purchases of a country's exports, the demand to hold liquid assets, labor demand, labor supply, etc.

Assume that x and y have been measured in m occasions, giving rise to a point cloud $(x_1, y_1), \dots, (x_m, y_m)$. Assume that some experts observe a linear trend, i.e., an apparent proportionality between increments of y and x . In order to predict the approximate value of y for a future observation x , we must fit a line of equation $y = ax + b$ to the given point cloud $(x_1, y_1), \dots, (x_m, y_m)$. The linear regression line of y on x is obtained by minimizing in some sense the residual vector

$$r = \begin{pmatrix} ax_1 + b - y_1 \\ \vdots \\ ax_m + b - y_m \end{pmatrix} \in \mathbb{R}^m$$

Fig. 1.5 In the ℓ_∞ linear regression, $f(a, b)$ is the length of the blue line segment



with respect to some norm. The most common measures of $r = (r_1, \dots, r_m)^T$ are its ℓ_2 (or Euclidean) norm $\|r\| := \sqrt{r_1^2 + \dots + r_m^2}$, its ℓ_1 (or Manhattan) norm $\|r\|_1 := |r_1| + \dots + |r_m|$, and its ℓ_∞ (or Chebyshev) norm $\|r\|_\infty := \max\{|r_1|, \dots, |r_m|\}$.

If one chooses the ℓ_∞ norm, the regression problem reads

$$P_1 : \text{Min } f(a, b) := \|r\|_\infty = \max\{|ax_1 + b - y_1|, \dots, |ax_m + b - y_m|\},$$

whose objective function

$$f(a, b) = \max\{|ax_1 + b - y_1|, \dots, |ax_m + b - y_m|\}$$

(see Fig. 1.5) is continuous but not differentiable, making unviable to solve P_1 through the identification of the critical points of f (points where the gradient is equal to zero). Recall that the maximum of two functions f and g which are continuous at a point is also continuous at that point as $\max\{f, g\} = \frac{f+g-|f-g|}{2}$.

Fortunately, P_1 is equivalent to

$$P_2 : \text{Min}_{(x, z) \in \mathbb{R}^{n+1}} z \\ \text{s.t.} \quad |ax_i + b - y_i| \leq z, \quad i = 1, \dots, m,$$

with variables a , b and z , which can be reformulated as a linear optimization problem by just replacing the i th constraint by two linear inequalities:

$$-z \leq ax_i + b - y_i \leq z.$$

1.1.5.6 ℓ_1 Linear Regression via Linear Optimization

Maintaining the notation above, the ℓ_1 linear regression is

$$P_1 : \text{Min} \|r\|_1 = |ax_1 + b - y_1| + \dots + |ax_m + b - y_m|,$$

whose objective function (see Fig. 1.6) is again continuous and nondifferentiable. Fortunately, this problem can be transformed into an equivalent linear optimization problem, since any linearly constrained nonlinear optimization problem whose objective function is a linear combination with positive coefficients of absolute values of affine functions can be reformulated as a linear optimization problem by introducing two new variables for each absolute value, as explained in Subsubsection 1.1.5.4.

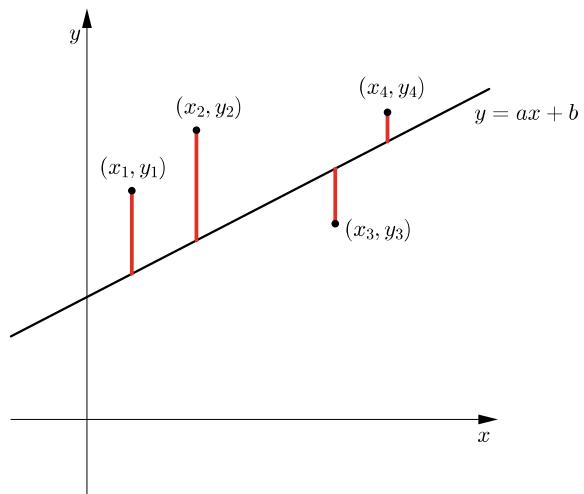
The affine functions that appear in the objective function of P_1 are of the form $(a, b) \mapsto ax_i + b - y_i$. Then, they can be expressed as $ax_i + b - y_i = u_i - v_i$, with $u_i \geq 0$ and $v_i \geq 0$, $i = 1, \dots, m$. The new objective function will be

$$f(a, b) = \|r\|_1 = \sum_{i=1}^m u_i + \sum_{i=1}^m v_i = 1_m^T(u + v),$$

where $1_m = (1, \dots, 1)^T$ denotes the vector of all ones, $u = (u_1, \dots, u_m)^T$ and $v = (v_1, \dots, v_m)^T$. We obtain in this way the equivalent model

$$\begin{aligned} P_2 : \quad & \text{Min}_{(u,v) \in \mathbb{R}^{2n}} 1_m^T(u + v) \\ \text{s.t.} \quad & u - v = \begin{pmatrix} ax_1 + b - y_1 \\ \vdots \\ ax_m + b - y_m \end{pmatrix}, \\ & u \geq 0_m, v \geq 0_m, \end{aligned}$$

Fig. 1.6 In the ℓ_1 linear regression, $f(a, b)$ is the sum of the lengths of the red line segments



whose $2(m + 1)$ unknowns are the unconstrained variables a, b (determining the aimed regression line), and the constrained variables u and v .

Example 1.7 It is easy to check that the ℓ_2 , ℓ_∞ and ℓ_1 regression lines of y over x for the tiny point cloud formed by the points $(-1, 0)$, $(0, 0)$ and $(1, 1)$ are $y = \frac{1}{2}x + \frac{1}{3}$, $y = \frac{1}{2}x + \frac{1}{4}$ and $y = \frac{1}{2}x + \frac{1}{2}$, respectively. The latter regression line is the hardest to be obtained as it requires to solve a linear optimization problem with eight unknowns.

1.2 Basic Calculus Tools

This section recalls basic concepts and calculus rules involving functions and matrices to be used in the sequel.

1.2.1 Complements of Differential Calculus

Definition 1.8 A *real-valued function of several variables* is a mapping from $A \subset \mathbb{R}^n$ on \mathbb{R} , in short, $f : A \rightarrow \mathbb{R}$. The set A is the *domain* of f , which is denoted by $\text{dom } f$. When f is given through a formula, $\text{dom } f$ is by default the set of vectors $x \in \mathbb{R}^n$ where $f(x)$ is well defined (i.e., it is a unique real number). The *graph* of f is the set

$$\text{gph } f = \left\{ \begin{pmatrix} x \\ f(x) \end{pmatrix} \in \mathbb{R}^{n+1} : x \in \text{dom } f \right\}.$$

When $n = 1$ and f is continuous, $\text{gph } f$ is a plane curve. When $n = 2$ and f is continuous, $\text{gph } f$ is a surface in the 3D space. In the latter case, we get a 2D representation of $\text{gph } f$ by appealing to the level curves. This type of representation is widely used in cartography.

Definition 1.9 The *level curve* k of a function $f : A \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ is the projection, on the plane $x_3 = 0$, of the intersection of $\text{gph } f$ with the horizontal plane $x_3 = k$, i.e., the set $\{x \in \text{dom } f : f(x) = k\}$.

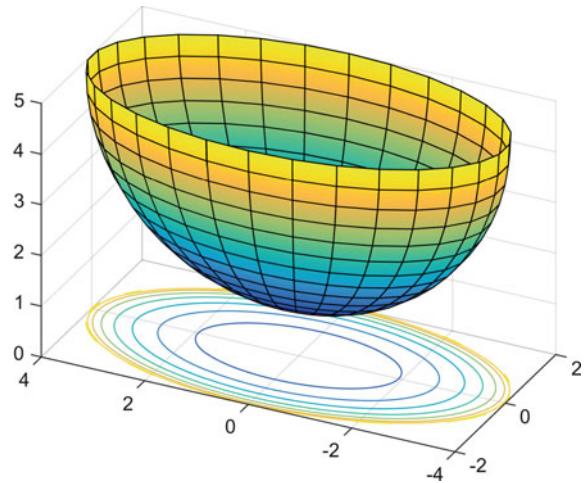
Example 1.10 Let $f(x) = 5 - \sqrt{16 - 4x_1^2 - x_2^2}$. We have

$$\text{dom } f = \{x \in \mathbb{R}^2 : 16 - 4x_1^2 - x_2^2 \geq 0\} = \left\{ x \in \mathbb{R}^2 : \frac{x_1^2}{4} + \frac{x_2^2}{16} \leq 1 \right\},$$

which is the region limited by the ellipse $\frac{x_1^2}{4} + \frac{x_2^2}{16} = 1$, while $\text{gph } f$ is the lower half of the ellipsoid

$$4x_1^2 + x_2^2 + (x_3 - 5)^2 = 16,$$

Fig. 1.7 Graph and level curves of
 $f(x) = 5 - \sqrt{16 - 4x_1^2 - x_2^2}$



whose center of symmetry is $(0, 0, 5)$, see Fig. 1.7. The level curve 5 is the ellipse $\frac{x_1^2}{4} + \frac{x_2^2}{16} = 1$ and the curve of level 1 is

$$\{x \in \text{dom } f : f(x) = 1\} = \{x \in \text{dom } f : 4x_1^2 + x_2^2 = 0\} = \{(0, 0)^T\}.$$

We now recall that, when $n = 1$, the *derivative* of f at $\bar{x} \in \text{int dom } f$ is

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

provided that this limit exists and is finite, that is,

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

i.e., the affine function $y = f(\bar{x}) + f'(\bar{x})(x - \bar{x})$ (whose graph is the tangent line to $\text{gph } f$ at $(\bar{x}, f(\bar{x}))$) is a good approximation of $y = f(x)$ close to \bar{x} . The linear function $h \mapsto f'(\bar{x})h$ is the *differential* of f at \bar{x} (its graph is the line through the origin which is parallel to the mentioned tangent). The slope of these lines, $f'(\bar{x})$, is the trigonometric tangent of the angle between the tangent line to $\text{gph } f$ at $(\bar{x}, f(\bar{x}))$ and the x -axis and represents the rate of variation of f when one moves from \bar{x} .

When $n > 1$ several concepts are rival candidates to extend the notion of derivative of f at $\bar{x} \in \text{int dom } f$. For the sake of simplicity, we assume $n = 2$.

Definition 1.11 The *partial derivative* of f with respect to x_1 at \bar{x} is the rate of variation of f when one moves from \bar{x} in the direction of the x_1 -axis. Precisely,

Fig. 1.8 Geometric interpretation of $\frac{\partial f}{\partial x_1}(1, 2)$

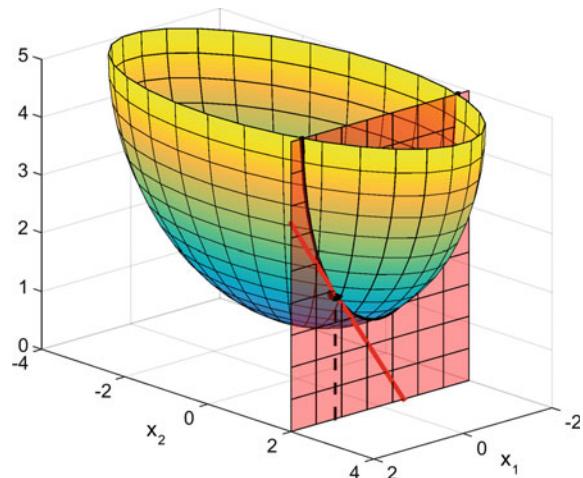
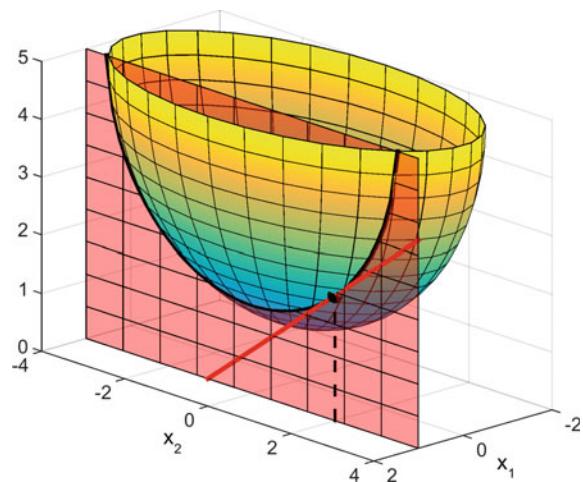


Fig. 1.9 Geometric interpretation of $\frac{\partial f}{\partial x_2}(1, 2)$



$$\frac{\partial f(\bar{x})}{\partial x_1} = \lim_{t \rightarrow 0} \frac{f(\bar{x}_1 + t, \bar{x}_2) - f(\bar{x}_1, \bar{x}_2)}{t},$$

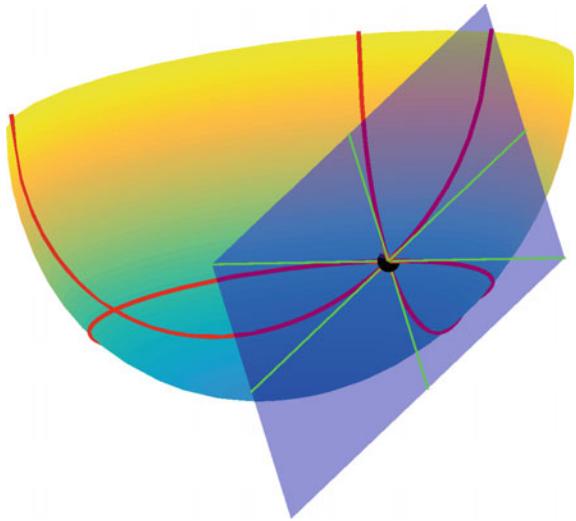
whose geometric meaning is nothing else than the slope of the tangent line to the intersection of $\text{gph } f$ with the vertical plane $x_2 = \bar{x}_2$.

Analogously, the *partial derivative* of f with respect to x_2 at \bar{x} is

$$\frac{\partial f(\bar{x})}{\partial x_2} = \lim_{t \rightarrow 0} \frac{f(\bar{x}_1, \bar{x}_2 + t) - f(\bar{x}_1, \bar{x}_2)}{t}.$$

The geometric interpretation of the partial derivatives of the function f in Example 1.10 is shown in Figs. 1.8 and 1.9.

Fig. 1.10 Tangent plane to $\text{gph } f$ at some given point



Definition 1.12 When both partial derivatives of the function f at \bar{x} exist, the *gradient* of f at \bar{x} is defined as the vector

$$\nabla f(\bar{x}) = \begin{pmatrix} \frac{\partial f(\bar{x})}{\partial x_1} \\ \frac{\partial f(\bar{x})}{\partial x_2} \end{pmatrix}.$$

Consider the tangent lines to the smooth curves contained in $\text{gph } f$ and passing through (\bar{x}) . In the case where all these tangent lines lie in a common plane, then that plane is called the *tangent plane* to $\text{gph } f$ at (\bar{x}) , see Fig. 1.10.

If both partial derivatives of f at \bar{x} exist, the curves obtained as intersections of $\text{gph } f$ with the planes $x_2 = \bar{x}_2$ and $x_1 = \bar{x}_1$ have parametric equations

$$\left. \begin{array}{l} x_1 = \bar{x}_1 + t \\ x_2 = \bar{x}_2 \\ x_3 = f(\bar{x}_1 + t, \bar{x}_2) \end{array} \right\} \quad \text{and} \quad \left. \begin{array}{l} x_1 = \bar{x}_1 \\ x_2 = \bar{x}_2 + t \\ x_3 = f(\bar{x}_1, \bar{x}_2 + t) \end{array} \right\},$$

whose tangent vectors at (\bar{x}) are given by the corresponding derivatives at $t = 0$, i.e., the vectors $\left(1, 0, \frac{\partial f(\bar{x})}{\partial x_1}\right)^T$ and $\left(0, 1, \frac{\partial f(\bar{x})}{\partial x_2}\right)^T$, which are orthogonal to the vector $\left(\begin{smallmatrix} \nabla f(\bar{x}) \\ -1 \end{smallmatrix}\right)$. Therefore, the vector $\left(\begin{smallmatrix} \nabla f(\bar{x}) \\ -1 \end{smallmatrix}\right)$ must be orthogonal to the tangent plane to $\text{gph } f$ at (\bar{x}) , provided that such a tangent plane does exist. That is, when the tangent plane exists, its equation is necessarily given by

Fig. 1.11 Tangent vectors to the curves (1.2.1) at $\begin{pmatrix} 1, 2, 5 - 2\sqrt{2} \end{pmatrix}^T$

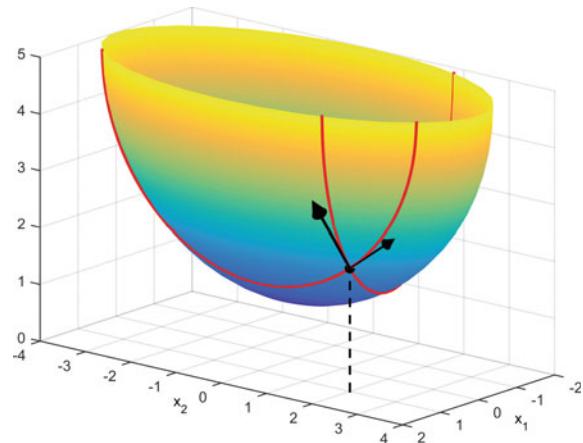
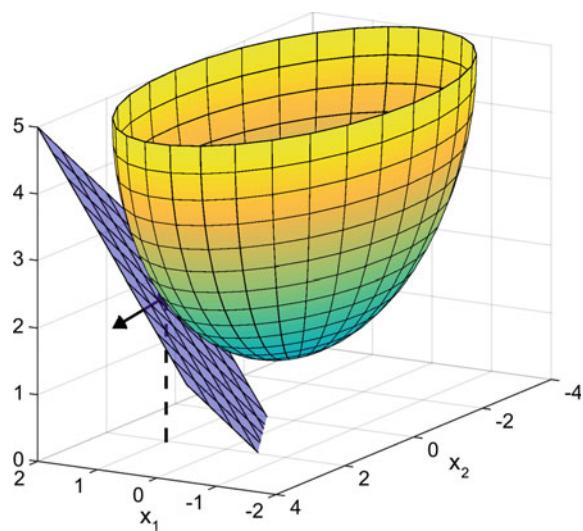


Fig. 1.12 Tangent plane and orthogonal vector to gph f at $\begin{pmatrix} 1, 2, 5 - 2\sqrt{2} \end{pmatrix}^T$ is $\begin{pmatrix} \nabla f(1, 2) \\ -1 \end{pmatrix}$



$$\begin{pmatrix} \nabla f(\bar{x}) \\ -1 \end{pmatrix}^T \begin{pmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \\ x_3 - f(\bar{x}_1, \bar{x}_2) \end{pmatrix} = 0,$$

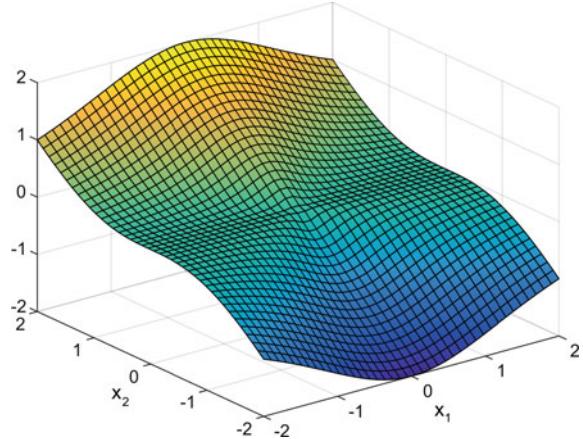
that is,

$$x_3 = f(\bar{x}_1, \bar{x}_2) + \nabla f(\bar{x})^T \begin{pmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \end{pmatrix}, \quad (1.5)$$

see Figs. 1.11 and 1.12.

The existence of all the partial derivatives is not enough to guarantee the existence of a tangent plane, as shown in Fig. 1.13. The definition of differentiability of f at \bar{x} requires that the affine function whose graph is given by (1.5), i.e., the tangent

Fig. 1.13 The function $f(x) = \frac{x_2^3}{x_1^2 + x_2^2}$ is not differentiable at the origin, despite the fact that all the partial derivatives exist



plane $x_3 = f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x})$, is a good approximation of f nearby \bar{x} . More precisely, the function f is said to be *differentiable* at $\bar{x} \in \text{int dom } f$ whenever its partial derivatives at \bar{x} exist and

$$\begin{aligned} & \lim_{x \rightarrow \bar{x}} \frac{f(x) - [f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x})]}{\|x - \bar{x}\|} \\ &= \lim_{h \rightarrow 0_2} \frac{f(\bar{x} + h) - [f(\bar{x}) + \nabla f(\bar{x})^T h]}{\|h\|} = 0 \end{aligned} \quad (1.6)$$

(we have done the change of variable $x = \bar{x} + h$). The *differential* of f at \bar{x} is the linear function $h \mapsto \nabla f(\bar{x})^T h$, whose graph is the plane passing through the origin which is parallel to the tangent plane to $\text{gph } f$ at $(\bar{x}, f(\bar{x}))$. Polynomial functions, among many others, are differentiable at any point of their domains.

When $u \in \mathbb{R}^2 \setminus \{0_2\}$ and f is differentiable at \bar{x} , taking $h = tu$ in (1.6), one gets

$$\lim_{t \rightarrow 0} \frac{f(\bar{x} + tu) - [f(\bar{x}) + t \nabla f(\bar{x})^T u]}{\|u\| |t|} = 0,$$

so we have that

$$\lim_{t \rightarrow 0} \frac{f(\bar{x} + tu) - [f(\bar{x}) + t \nabla f(\bar{x})^T u]}{t} = 0. \quad (1.7)$$

From (1.7), it follows

$$\lim_{t \rightarrow 0} \frac{f(\bar{x} + tu) - f(\bar{x})}{t} = \nabla f(\bar{x})^T u. \quad (1.8)$$

When the first member of (1.8) exists and is finite, it is called the *directional derivative* of f at \bar{x} in the direction of u , and is denoted by $f'(\bar{x}; u)$. It represents the rate of

change of f when one moves from \bar{x} in the direction of u , provided that $\|u\| = 1$, see Fig. 1.14. Hence, when f is differentiable at \bar{x} ,

$$f'(\bar{x}; u) = \nabla f(\bar{x})^T u. \quad (1.9)$$

We have seen that, when f is differentiable at \bar{x} , the tangent plane to $\text{gph } f$ at $(\begin{smallmatrix} \bar{x} \\ f(\bar{x}) \end{smallmatrix})$ is orthogonal to $(\begin{smallmatrix} \nabla f(\bar{x}) \\ -1 \end{smallmatrix})$. Thus, the latter vector is orthogonal to any tangent vector at $(\begin{smallmatrix} \bar{x} \\ f(\bar{x}) \end{smallmatrix})$ of the curves contained in the surface $\text{gph } f$. In particular, it must be orthogonal to the tangent vector to the curve resulting from the intersection of $\text{gph } f$ with the horizontal plane passing through $(\begin{smallmatrix} \bar{x} \\ f(\bar{x}) \end{smallmatrix})$, whose equation is $x_3 = f(\bar{x})$. Such an horizontal tangent vector is of the form $(\begin{smallmatrix} v \\ 0 \end{smallmatrix})$, where v is tangent to the curve of level $f(\bar{x})$, i.e., $\{x \in \text{dom } f : f(x) = f(\bar{x})\}$. We thus have

$$\left(\begin{array}{c} \nabla f(\bar{x}) \\ -1 \end{array} \right)^T \left(\begin{array}{c} v \\ 0 \end{array} \right) = \nabla f(\bar{x})^T v = 0.$$

Hence, $\nabla f(\bar{x})$ is orthogonal to the level curve which contains \bar{x} . On the other hand, if $\|u\| = 1$ and we denote by θ the angle between u and $\nabla f(\bar{x})$, by (1.9) we have

$$f'(\bar{x}; u) = \nabla f(\bar{x})^T u = \|\nabla f(\bar{x})\| \|u\| \cos \theta = \|\nabla f(\bar{x})\| \cos \theta,$$

whose greatest (lowest) value is attained whenever $\cos \theta = 1$ ($\cos \theta = -1$), i.e., $\theta = 0$ ($\theta = \pi$, respectively). Therefore, the directional derivative of f at \bar{x} attains its greatest (lowest) possible value whenever $u = \frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|}$ ($u = -\frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|}$, respectively). In conclusion, $\nabla f(\bar{x})$ turns out to be orthogonal to the level curve of f through \bar{x} and its direction is the one providing its greatest rate of growth.

Recall that f is continuously differentiable (f is \mathcal{C}^1 , in brief) at \bar{x} when all the first partial derivatives of f exist in some neighborhood of \bar{x} and are continuous at \bar{x} . Similarly, f is twice continuously differentiable (f is \mathcal{C}^2) at \bar{x} if, additionally, the second partial derivatives of f exist in some neighborhood of \bar{x} and are continuous at \bar{x} . When f is \mathcal{C}^i at all the points of some set $U \subset \mathbb{R}^n$, then we write $f \in \mathcal{C}^i(U)$, $i = 1, 2$.

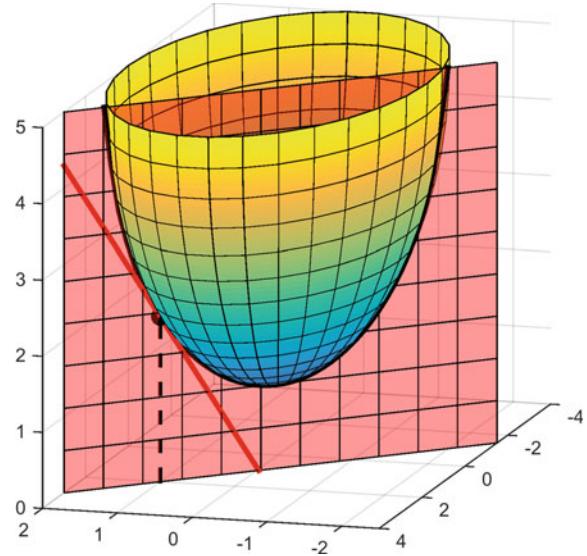
Definition 1.13 The *Hessian matrix* of f at a point x where f is twice differentiable is given by

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial^2 x_1} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 f(x)}{\partial^2 x_n} \end{bmatrix}.$$

When f is \mathcal{C}^2 at \bar{x} , i.e., when f is \mathcal{C}^1 on some neighborhood of \bar{x} and $\frac{\partial^2 f(x)}{\partial x_j \partial x_i}$ is continuous at \bar{x} , $i, j = 1, \dots, n$, then $\nabla^2 f(\bar{x})$ is symmetric (Schwarz's theorem).

In Part II, we will use *asymptotic* (or *Landau*) notation to describe the limiting behavior of a function when the argument tends to zero. Given two functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}$, we write $f(x) = o(g(x))$ whenever

Fig. 1.14 Geometric interpretation of
 $f' \left((1, 2)^T; \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)^T \right)$



$$\lim_{\substack{g(x) \rightarrow 0 \\ g(x) \neq 0}} \frac{f(x)}{g(x)} = 0.$$

For instance, given a function $h : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and a positive integer q , we write $h(x) = o(\|x\|^q)$ whenever

$$\lim_{x \rightarrow 0} \frac{h(x)}{\|x\|^q} = 0.$$

According to Taylor's theorem, if f is \mathcal{C}^2 on some open set containing the segment $[\bar{x}, x]$, with $x \neq \bar{x}$, then there exists a point $z_x \in]\bar{x}, x[$ such that

$$f(x) = f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2} (x - \bar{x})^T \nabla^2 f(z_x) (x - \bar{x}).$$

Assume that $f \in \mathcal{C}^2(U)$, where $U = \bar{x} + \rho \mathbb{B}$, with $\mathbb{B} := \{x \in \mathbb{R}^n : \|x\| \leq 1\}$ (the Euclidean closed unit ball). Then, for any $p \in \rho \mathbb{B}$, there exists $z_p \in]\bar{x}, \bar{x} + p[$ such that

$$f(\bar{x} + p) = f(\bar{x}) + \nabla f(\bar{x})^T p + \frac{1}{2} p^T \nabla^2 f(z_p) p.$$

Define $r(p) := \frac{1}{2} p^T (\nabla^2 f(z_p) - \nabla^2 f(\bar{x})) p$. Since $f \in \mathcal{C}^2(U)$, for all $\varepsilon > 0$ there exists $\delta > 0$, with $\delta < \rho$, such that $\|z - \bar{x}\| < \delta$ entails

$$\left| \frac{\partial^2 f(z)}{\partial x_i \partial x_j} - \frac{\partial^2 f(\bar{x})}{\partial x_i \partial x_j} \right| \leq \frac{2\varepsilon}{n} \quad \text{for all } i, j = 1, \dots, n.$$

Let $p \in \delta\mathbb{B}$. Then, since $\|z_p - \bar{x}\| < \delta$, we have

$$\begin{aligned} |r(p)| &\leq \frac{\varepsilon}{n} \sum_{i=1}^n \sum_{j=1}^n |p_i p_j| = \frac{\varepsilon}{n} \left(\sum_{i=1}^n |p_i| \right)^2 = \frac{\varepsilon}{n} ((|p_1|, \dots, |p_n|) 1_n)^2 \\ &\leq \frac{\varepsilon}{n} \|p\|^2 \|1_n\|^2 = \varepsilon \|p\|^2, \end{aligned}$$

where we have applied the Cauchy–Schwarz inequality to get the second inequality. Hence,

$$\frac{|r(p)|}{\|p\|^2} \leq \varepsilon.$$

The fulfillment of the latter inequality for all $p \in \delta\mathbb{B}$ and $\varepsilon > 0$ implies that $r(p) = o(\|p\|^2)$. Hence, we can write the second-order Taylor formula as

$$f(\bar{x} + p) = f(\bar{x}) + \nabla f(\bar{x})^T p + \frac{1}{2} p^T \nabla^2 f(\bar{x}) p + o(\|p\|^2). \quad (1.10)$$

The quadratic function

$$x \mapsto f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \frac{1}{2} (x - \bar{x})^T \nabla^2 f(\bar{x}) (x - \bar{x})$$

is the second-order approximation of f close to \bar{x} .

Example 1.14 In Example 1.10, we have

$$\nabla f(x) = \begin{pmatrix} \frac{4x_1}{\sqrt{16-4x_1^2-x_2^2}} \\ \frac{x_2}{\sqrt{16-4x_1^2-x_2^2}} \end{pmatrix} \text{ and } \nabla^2 f(x) = \begin{bmatrix} \frac{64-4x_2^2}{(16-4x_1^2-x_2^2)^{\frac{3}{2}}} & \frac{4x_1 x_2}{(16-4x_1^2-x_2^2)^{\frac{3}{2}}} \\ \frac{4x_1 x_2}{(16-4x_1^2-x_2^2)^{\frac{3}{2}}} & \frac{16-x_1^2}{(16-4x_1^2-x_2^2)^{\frac{3}{2}}} \end{bmatrix},$$

so we have that the second-order approximation of f close to $(0, 0)$ is

$$\begin{aligned} q(x) &= f(0, 0) + \nabla f(0, 0)^T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \frac{1}{2} (x_1, x_2) \nabla^2 f(0, 0) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= 1 + \frac{1}{2} x_1^2 + \frac{1}{8} x_2^2. \end{aligned}$$

We have considered up to now in this subsection scalar functions.

Definition 1.15 A *multivariate function* (or *vector function*) f is a mapping from a set $A \subset \mathbb{R}^n$ on \mathbb{R}^m . We write $f = (f_1, \dots, f_m) : A \rightarrow \mathbb{R}^m$, where the i -th component f_i of f is the real-valued function that associates to each $x \in A$ the i -th coordinate of $f(x) \in \mathbb{R}^m$. The set A is the *domain* of f , denoted by $\text{dom } f$ (by default, $\text{dom } f$ is the set of points $x \in \mathbb{R}^n$ where $f(x)$ is well defined).

Definition 1.16 The *matrix of gradients* of $f : A \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ at a point $\bar{x} \in \text{int dom } f$ where all the components of f are differentiable is the $n \times m$ matrix whose columns are the gradients of the functions f_i , $i = 1, 2, \dots, m$, i.e.,

$$\nabla f(x) = [\nabla f_1(x) \mid \dots \mid \nabla f_m(x)] = \begin{bmatrix} \frac{\partial f_1(\bar{x})}{\partial x_1} & \dots & \frac{\partial f_m(\bar{x})}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial f_1(\bar{x})}{\partial x_n} & \dots & \frac{\partial f_m(\bar{x})}{\partial x_n} \end{bmatrix}.$$

The *Jacobian matrix* of f , denoted by $J_f(x)$ is the transpose of the gradient matrix, i.e.,

$$J_f(x) = \nabla f(x)^T = \begin{bmatrix} \nabla f_1(x)^T \\ \vdots \\ \nabla f_m(x)^T \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(\bar{x})}{\partial x_1} & \dots & \frac{\partial f_1(\bar{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\bar{x})}{\partial x_1} & \dots & \frac{\partial f_m(\bar{x})}{\partial x_n} \end{bmatrix}.$$

When $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is a linear mapping of the form $f(x) = Mx$, where M is a given $m \times n$ matrix, we have $J_f(x) = M$ and $\nabla f(x) = M^T$ for all $x \in \mathbb{R}^n$. When f is a real-valued function of a single variable, i.e., $m = 1$, differentiable at \bar{x} , $\nabla f(\bar{x})$ is the 1×1 matrix $[f'(\bar{x})]$, which thus coincides with $J_f(\bar{x})$. For the latter type of functions, the so-called chain rule for real-valued functions of one variable asserts that, if f can be composed with g , f is differentiable at \bar{x} and g is differentiable at $f(\bar{x})$, then the composite function $g \circ f$ is differentiable at \bar{x} with derivative $(g \circ f)'(\bar{x}) = g'(f(\bar{x}))f'(\bar{x})$.

The multivariate counterpart of the above chain rule of elementary calculus is the following *multivariate chain rule*. Let $f : A \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $g : B \subset \mathbb{R}^m \rightarrow \mathbb{R}^p$ be such that $f(A) \subset B$, and let $\bar{x} \in A$ be such that f is differentiable at \bar{x} and g is differentiable at $f(\bar{x})$. Then, $g \circ f$ is differentiable at \bar{x} and its Jacobian matrix is the product of the Jacobian matrices of f and g at \bar{x} and $f(\bar{x})$, respectively, i.e.,

$$J_{g \circ f}(\bar{x}) = J_g(f(\bar{x}))J_f(\bar{x}),$$

while the order is reversed for its gradient:

$$\nabla(g \circ f)(\bar{x}) = \nabla f(\bar{x})\nabla g(f(\bar{x})).$$

Observe that, if $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $g : \mathbb{R}^m \rightarrow \mathbb{R}^p$ are the linear functions $f(x) = Mx$ and $g(y) = Ny$, with M and N given $m \times n$ and $p \times m$ matrices, then the multivariate chain rule yields $J_{g \circ f}(\bar{x}) = NM$ (as expected, since $(g \circ f)(x) = (NM)x$).

The multivariate chain rule is frequently used in this book to obtain the gradient of $g \circ f$ when either f is an affine function of a real variable and g is real-valued (as in Section 1.3 and Chapter 2), or when f is a multivariate function and g is the real-valued function $g(y) = \frac{1}{2}\|y\|^2$ (as in Chapters 5 and 6):

- Let $f : \mathbb{R} \rightarrow \mathbb{R}^m$ be the mapping $f(x) = a + xd$, with $a, d \in \mathbb{R}^m$ and $g : B \subset \mathbb{R}^m \rightarrow \mathbb{R}$. Then,

$$\nabla(g \circ f)(\bar{x}) = \nabla f(\bar{x}) \nabla g(f(\bar{x})) = d^T \nabla g(f(\bar{x})).$$

- Let $f : A \subset \mathbb{R}^n \rightarrow \mathbb{R}^m$ and $g : \mathbb{R}^m \rightarrow \mathbb{R}$, with $g(y) = \frac{1}{2}\|y\|^2$. Then,

$$\nabla(g \circ f)(\bar{x}) = \nabla f(\bar{x}) \nabla g(f(\bar{x})) = \nabla f(\bar{x}) f(\bar{x}).$$

1.2.2 Complements of Matrix Calculus

The most useful generalization of the concept of nonnegative real number is the concept of semidefinite matrix, which frequently arises in optimal design problems (engineering, see [57]). We denote by \mathcal{S}_n the family of symmetric $n \times n$ matrices.

Definition 1.17 The matrix $A \in \mathcal{S}_n$ is *positive semidefinite* if $x^T Ax \geq 0$ for all $x \in \mathbb{R}^n$. In particular, $A \in \mathcal{S}_n$ is *positive definite* whenever $x^T Ax > 0$ for all $x \in \mathbb{R}^n \setminus \{0_n\}$. Analogously, $A \in \mathcal{S}_n$ is *negative semidefinite (definite)* when $-A$ is positive semidefinite (definite, respectively). Finally, $A \in \mathcal{S}_n$ is *indefinite* when it is neither positive semidefinite nor negative semidefinite.

Definition 1.18 The *Gram matrix* of the $n \times m$ matrix M is $MM^T \in \mathcal{S}_n$.

Gram matrices appear in many closed formulas that provide optimal solutions.

Proposition 1.19 (Sign of Gram matrix) *The Gram matrix MM^T of any $n \times m$ matrix M is positive semidefinite. Moreover, it is positive definite if and only if M is row-rank complete.*

Proof Let $G := MM^T$. Given $x \in \mathbb{R}^n$, we have

$$x^T G x = x^T M M^T x = \|M^T x\|^2 \geq 0,$$

so G is always positive semidefinite. Two cases may arise:

When M is row-rank complete and $x \in \mathbb{R}^n \setminus \{0_n\}$, then $M^T x \neq 0_m$ as this is a nontrivial linear combination of linearly independent vectors (the columns of M^T), so we have that $x^T G x = \|M^T x\|^2 > 0$ and G is positive definite.

When M is not row-rank complete, there exists $x \in \mathbb{R}^n \setminus \{0_n\}$ such that $M^T x = 0_m$. Then, $x^T G x = \|M^T x\|^2 = 0$ and G is not positive definite. \square

Remark 1.20 As a consequence of Proposition 1.19, the Gram matrix of a nonsingular square matrix is positive definite.

A square matrix A is singular if and only if zero is an eigenvalue of A . It is easy to show that, if A is nonsingular, then A^{-1} has the same eigenvectors as A , while the product of the corresponding eigenvalues is 1.

It is well known that, given $A \in \mathcal{S}_n$, there exists an orthogonal matrix U such that

$$U^T A U = \text{diag}(\lambda_1, \dots, \lambda_n), \quad (1.11)$$

where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of A (repeated as many times as their corresponding multiplicity orders). Defining $y := U^T x$, one has

$$x^T A x = \sum_{i=1}^n \lambda_i y_i^2,$$

with $x = 0_n$ if and only if $y = 0_n$. Thus, A is positive semidefinite (positive definite) if and only if $\lambda_i \geq 0$ for all $i = 1, \dots, n$ ($\lambda_i > 0$ for all $i = 1, \dots, n$, respectively). Therefore, A is positive definite if and only if it is positive semidefinite and $\det A > 0$ while it is indefinite when A has eigenvalues of different signs.

We can assume without loss of generality that $\lambda_{\min} := \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n := \lambda_{\max}$ (reorder the eigenvectors of A and the columns of U if necessary). Since

$$\sum_{i=1}^n y_i^2 = y^T y = (x^T U)(U^T x) = x^T x = \|x\|^2,$$

we have

$$\lambda_{\min} \|x\|^2 \leq x^T A x = \sum_{i=1}^n \lambda_i y_i^2 \leq \lambda_{\max} \|x\|^2. \quad (1.12)$$

Thanks to (1.11) and the orthogonality of U , we have $A = U \text{diag}(\lambda_1, \dots, \lambda_n) U^T$. Denoting by u_1, \dots, u_n the columns of U (a basis of \mathbb{R}^n formed by eigenvectors of A), one gets the so-called *spectral decomposition* of A , which expresses A as a sum of rank 1 symmetric matrices:

$$A = \sum_{i=1}^n [0_n \mid \dots \mid u_i \mid \dots \mid 0_n] \begin{bmatrix} \lambda_1 u_1^T \\ \vdots \\ \lambda_n u_n^T \end{bmatrix} = \sum_{i=1}^n \lambda_i u_i u_i^T.$$

Therefore, any positive semidefinite matrix is a nonnegative linear combination of rank 1 symmetric matrices (which are positive semidefinite too).

We now show the existence and uniqueness of the *square root* for any positive semidefinite matrix A , i.e., a positive semidefinite matrix $A^{\frac{1}{2}}$ such that

$$\left(A^{\frac{1}{2}}\right)^2 = A.$$

We first consider the simple case where $n = 1$. Obviously, $\sqrt{0} = 0$ because $0^2 = 0$ and \mathbb{R} is a field (so we have that $x^2 \neq 0$ for all $x \neq 0$), while Bolzano's theorem applied to the function $p(x) := x^2 - \lambda$ (for which $p(0) < 0$ and $\lim_{x \rightarrow \pm\infty} p(x) = +\infty$) shows the existence of two numbers of different sign, $\pm\sqrt{\lambda}$, such that $x^2 = \lambda$. The uniqueness of $\sqrt{\lambda}$ comes from the fundamental theorem of algebra (p has exactly two complex zeros, and thus, it has at most two real zeros). We now consider an $n \times n$, $n \geq 2$, positive semidefinite matrix A . If $A = U \text{diag}(\lambda_1, \dots, \lambda_n)U^T$, with U being an orthogonal matrix, it is easy to see that $S := U \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n})U^T$ is a positive semidefinite matrix satisfying $S^2 = A$. To prove the uniqueness it is worth observing that, if λ is an eigenvalue of a matrix $M \in \mathcal{S}_n$ associated with an eigenvector v and W is an orthogonal matrix, one has

$$(WMW^T)Wv = WMv = \lambda Wv, \quad (1.13)$$

that is, λ is an eigenvalue of WMW^T with corresponding eigenvector Wv . Assume now that S is a positive semidefinite matrix such that $S^2 = A$, with

$$U^T AU = \text{diag}(\lambda_1, \dots, \lambda_n) \quad \text{and} \quad V^T SV = \text{diag}(\alpha_1, \dots, \alpha_n),$$

where $\lambda_1, \dots, \lambda_n, \alpha_1, \dots, \alpha_n \in \mathbb{R}_+$ (the set of nonnegative real numbers), and the matrices U and V are orthogonal. We thus have

$$\text{diag}(\alpha_1^2, \dots, \alpha_n^2) = (V^T U) \text{diag}(\lambda_1, \dots, \lambda_n) (U^T V), \quad (1.14)$$

where $V^T U$ is orthogonal and, by (1.13), the diagonal matrices in (1.14) have the same eigenvalues, i.e., $\alpha_i = \sqrt{\lambda_i}$, $i = 1, \dots, n$.

In particular, when A is positive definite,

$$\left(A^{\frac{1}{2}}\right)^{-1} = U \text{diag}\left(\frac{1}{\sqrt{\lambda_1}}, \dots, \frac{1}{\sqrt{\lambda_n}}\right) U^T = (A^{-1})^{\frac{1}{2}},$$

so it is also possible to define $A^{-\frac{1}{2}}$ as the unique $n \times n$ positive definite matrix such that $\left(A^{-\frac{1}{2}}\right)^2 = A^{-1}$.

The main drawback of the classification criterion for symmetric matrices based on the sign of the eigenvalues is that it requires the computation of all zeros of the *characteristic polynomial* $\det(A - \lambda I_n)$, where I_n is the $n \times n$ identity matrix, by means of some numerical method (as e.g. Newton's method). Fortunately, there are other ways to classify a matrix $A \in \mathcal{S}_n$ as either positive semidefinite, negative semidefinite, or indefinite. In particular, we show next three ways of doing this: computing the principal minors of A , calculating its invariants, and solving an optimization problem associated to A . It is also possible to check in a similar manner whether a given positive (negative) semidefinite matrix is positive (negative) definite.

Denote by A_k the $k \times k$ submatrix of A which results of taking the first k rows and columns; its determinant, $\Delta_k = \det A_k$, is the k -th *director principal minor*

of A . It is well known that A is positive definite if and only if $\Delta_k > 0$, $k = 1, \dots, n$. Checking the positive semidefiniteness of A requires more computations, as A is positive semidefinite if and only if all *principal minors* of A (determinants of the submatrices obtained from A by elimination of rows and columns of the same index) are all nonnegative. The next example shows that the director principal minors do not serve to certify that a given symmetric matrix is positive semidefinite.

Example 1.21 The director principal minors of

$$A = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & \frac{1}{2} \end{bmatrix}$$

are all nonnegative, but neither A nor B are positive semidefinite (take the vectors $(0, 1)^T$ and $(1, 1, -2)^T$, or consider their respective eigenvalues 0 and -1 , and 0 and $\frac{5 \pm \sqrt{41}}{4}$).

The characteristic polynomial of the matrix A can be written as

$$\det(A - \lambda I_n) = (-\lambda)^n + i_1(A)(-\lambda)^{n-1} + \dots + i_n(A),$$

where the coefficients of the powers of $-\lambda$ are called *invariants* of A (e.g., when $n = 2$, the invariants are nothing else than the trace and the determinant of A). For instance, the characteristic polynomial of the matrix B in Example 1.21 is $-\lambda^3 + \frac{5}{2}\lambda^2 + \lambda$, so we have that its invariants are $i_1(B) = \frac{5}{2}$, $i_2(B) = -1$ and $i_3(B) = 0$.

It is known that A is positive semidefinite (positive definite) if and only if $i_k(A) \geq 0$, $k = 1, \dots, n$ ($i_k(A) > 0$, $k = 1, \dots, n$, respectively).

Example 1.22 The characteristic polynomial of

$$A = \begin{bmatrix} 2 & -4 & 0 \\ -4 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

is

$$\det(A - \lambda I_3) = -\lambda^3 + 6\lambda^2 + 4\lambda - 24 = (-\lambda)^3 + 6(-\lambda)^2 - 4(-\lambda) - 24.$$

As $i_2(A) = -4 < 0$, A is not even positive semidefinite. In fact, its eigenvalues are 6 and ± 2 , so A is indefinite.

To conclude this part devoted to classify symmetric matrices by their sign, let us associate with $A \in \mathcal{S}_n$ the unconstrained optimization problem

$$P_A : \text{Min } q_A(x) := x^T A x.$$

It is immediate to show that A is positive semidefinite if and only if $v(P_A) = 0$. In that case, A is positive definite if and only if $\det A \neq 0$, in which case the unique optimal solution of P_A is 0_n .

We finish this section by recalling the notion of norm of a matrix and introducing the concept of condition number of a square matrix, which will be widely used in Part II. The *spectral norm* of a given $m \times n$ matrix A is

$$\|A\| := \max_{\|x\|=1} \|Ax\|.$$

If $x \in \mathbb{R}^n \setminus \{0_n\}$, by the Cauchy–Schwarz inequality,

$$\|Ax\| = \|x\| \left\| A \left(\frac{x}{\|x\|} \right) \right\| \leq \|A\| \|x\|, \quad (1.15)$$

where the equality holds whenever $x = 0_n$. If $A = [a_{ij}]$, and $\mu := \max_{i,j} |a_{ij}|$, it is easy to show that

$$\|A\| \leq \mu n \sqrt{n}, \quad (1.16)$$

which provides an upper bound for $\|A\|$ in terms of the entries of A .

If A is an $m \times n$ matrix, B is an $n \times p$ matrix, and y is a unitary vector in \mathbb{R}^p , one has

$$\|(AB)y\| = \|A(By)\| \leq \|A\| \|By\| \leq \|A\| \|B\| \|y\| = \|A\| \|B\|,$$

so we have that $\|AB\| \leq \|A\| \|B\|$.

It is known that the spectral norm of A can also be expressed as

$$\|A\| = \sqrt{\rho(A^T A)},$$

where $\rho(A^T A)$ is the so-called *spectral radius* of $A^T A$ (the Gram matrix of A^T), whose value is the greatest eigenvalue of $A^T A$ (see Exercise 6.14).

The rest of this section deals with square matrices. First we give a quadratic version of the inequality (1.15) to be used later. Let A be an $n \times n$ matrix and $x \in \mathbb{R}^n \setminus \{0_n\}$. Then, appealing again to the Cauchy–Schwarz inequality, we have

$$x^T A x = \|x\| x^T \left(A \frac{x}{\|x\|} \right) \leq \|x\|^2 \left\| A \frac{x}{\|x\|} \right\| \leq \|A\| \|x\|^2, \quad (1.17)$$

where the equality holds whenever $x = 0_n$.

Definition 1.23 The *condition number* of an $n \times n$ matrix A is defined as

$$\text{cond}(A) := \|A\| \|A^{-1}\|,$$

when A is nonsingular, and $\text{cond}(A) := +\infty$ when A is singular.

The following properties of the condition number of a nonsingular matrix can be easily shown:

- $\text{cond}(A) \geq 1$, as $\|A\|\|A^{-1}\| \geq \|AA^{-1}\| = \|I_n\| = 1$;
- $\text{cond}(A) = \text{cond}(A^{-1})$;
- $\text{cond}(\lambda A) = \text{cond}(A)$, for all $\lambda \neq 0$.

In the particular case where A is symmetric, with eigenvalues $\lambda_{\min} = \lambda_1 \leq \dots \leq \lambda_n = \lambda_{\max}$, it is known that

$$\|A\| = \sqrt{\rho(A^2)} = \sqrt{\max\{|\lambda_{\min}|^2, |\lambda_{\max}|^2\}} = \max\{|\lambda_{\min}|, |\lambda_{\max}|\}.$$

When, additionally, A is positive definite, $\|A\| = \lambda_{\max}$ and its condition number is

$$\text{cond}(A) = \|A\|\|A^{-1}\| = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

Square matrices whose condition number is close to 1 are said to be *well-conditioned*. On the contrary, when the condition number is large, the matrix is said to be *ill-conditioned*.

The condition number of a square matrix A (or a linear system whose left-hand side data is represented by A) is a measure of its stability or sensitivity to numerical operations. In other words, we should not trust computational results involving ill-conditioned matrices. Consider, e.g., a linear system $Ax = b$, where A is an $n \times n$ nonsingular matrix, and let \bar{x} be its solution. Perturbing A to \tilde{A} and b to \tilde{b} , and denoting by \tilde{x} the solution of the perturbed system $\tilde{A}x = \tilde{b}$ (assuming that \tilde{A} is still nonsingular), one has

$$\frac{\|\bar{x} - \tilde{x}\|}{\|\bar{x}\|} \approx \text{cond}(A) \left(\frac{\|A - \tilde{A}\|}{\|A\|} + \frac{\|b - \tilde{b}\|}{\|b\|} \right)$$

(see, e.g., [45, Section 2.7, pp. 80–81]). An example of an ill-conditioned system is given next.

Example 1.24 The (exact) solution of

$$\begin{bmatrix} 1.00001 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2.00001 \\ 2 \end{bmatrix}$$

is $\bar{x} = (1, 1)^T$. However, if we replace 2.00001 by 2.00002 in the right-hand side, the solution of the perturbed system abruptly changes to $\tilde{x} = (0, 2)^T$. The condition number of the matrix

$$A = \begin{bmatrix} 1.00001 & 1 \\ 1 & 1 \end{bmatrix}$$

can be directly computed from the definition of the spectral norm as

$$\text{cond}(A) = \|A\| \|A^{-1}\| = 2 \times (2 \times 10^5) = 4 \times 10^5,$$

or, exploiting the fact that A is positive definite, from its eigenvalues, as

$$\text{cond}(A) = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{2}{5 \times 10^{-6}} = 4 \times 10^5.$$

The missing proofs in this section can be found, e.g., in either the monograph [58] or the review paper [57].

1.3 Optimality Conditions at Interior Points

The optimization problem (1.1) can be expressed as

$$\begin{aligned} P : \text{Min } & f(x) \\ \text{s.t. } & x \in F, \end{aligned}$$

where $F \subset \mathbb{R}^n$ represents the feasible set.

The Fermat principle establishes that any local minimum of f at the interior of F is a *critical point* (also called *stationary point*), that is, a point where the gradient vanishes. A *saddle point* is a critical point but not a local extremum.

Theorem 1.25 (First-order optimality conditions) *Let $\bar{x} \in \text{int } F$ and assume that $f \in \mathcal{C}^1(U)$ for some neighborhood U of \bar{x} . The following statements hold true:*

- (i) *If \bar{x} is a local minimum of f , then $\nabla f(\bar{x}) = 0_n$.*
- (ii) *If $\nabla f(x)^T(x - \bar{x}) > 0$ for all $x \in U \setminus \{\bar{x}\}$, then f has a strict local minimum at \bar{x} .*

Proof (i) Assume that $\nabla f(\bar{x}) \neq 0_n$. Let $u := -\nabla f(\bar{x})$. Then,

$$\lim_{t \searrow 0} \frac{f(\bar{x} + tu) - f(\bar{x})}{t} = f'(\bar{x}; u) = \nabla f(\bar{x})^T u = -\|\nabla f(\bar{x})\|^2 < 0,$$

so we have that $f(\bar{x} + tu) - f(\bar{x}) < 0$ for all $t > 0$ sufficiently small. Hence, \bar{x} is not a local minimum of f .

(ii) We can assume without loss of generality that $U = \bar{x} + \varepsilon \mathbb{B}$, with $\varepsilon > 0$. Given an arbitrary $x \in U \setminus \{\bar{x}\}$, consider the function $\Phi : [0, 1] \rightarrow \mathbb{R}$ defined as $\Phi(t) = f(\bar{x} + t(x - \bar{x}))$ for all $t \in [0, 1]$. Since f is continuous and differentiable on U , Φ is continuous on $[0, 1]$, differentiable on $]0, 1[$, and

$$\Phi'(t) = \nabla f(\bar{x} + t(x - \bar{x}))^T(x - \bar{x})$$

for all $t \in]0, 1[$, because of the chain rule. By the mean value theorem, there exists $t_0 \in]0, 1[$ such that $\Phi(1) - \Phi(0) = \Phi'(t_0)$ or, equivalently,

$$f(x) - f(\bar{x}) = \nabla f(\bar{x} + t_0(x - \bar{x}))^T(x - \bar{x}) = \frac{1}{t_0} \nabla f(y)^T(y - \bar{x}), \quad (1.18)$$

where $y := \bar{x} + t_0(x - \bar{x})$. As $\|y - \bar{x}\| = t_0\|x - \bar{x}\| < \|x - \bar{x}\| < \varepsilon$, one has $y \in U \setminus \{\bar{x}\}$ and, by assumption, $\nabla f(y)^T(y - \bar{x}) > 0$. From this inequality, together with (1.18), we get that $f(x) > f(\bar{x})$ for all $x \in U \setminus \{\bar{x}\}$, i.e., \bar{x} is a strict local minimum of f . \square

Observe that the proof of (i) only requires the differentiability of f at \bar{x} . Notice also that the sufficient condition in (ii) states that \bar{x} is a strict local minimum of f whenever the directional derivative at any point x in some neighborhood of \bar{x} in the direction from x to \bar{x} are all negative. For example, according to (i), the unique candidate to be a local minimum of $f(x) = x^4$ is 0; since $f'(x)x = 4x^3 > 0$ for all $x \neq 0$, then 0 is a strict local minimum. Similarly, 0_2 is the unique candidate to be local minimum of $f(x) = \|x\|^2$; it is actually a strict local minimum as $\nabla f(x)^T x = 2\|x\|^2 > 0$ for all $x \neq 0_2$.

When we know for sure that a global minimum of P does exist, we can identify it by comparing the value of f at the boundary of F with the value of f at the critical points, provided that the set of critical points is finite (a simple function as $f(x) = x^2 \sin^2 x$ has countable many critical points). The Weierstrass theorem guarantees the existence of a global minimum of f on $F \subset \mathbb{R}^n$ when F is compact (that is, when F is a closed bounded set). A weaker condition (coercivity) will be introduced in the next section.

Example 1.26 The candidates to local minima for the function of Example 1.10, $f(x) = 5 - \sqrt{16 - 4x_1^2 - x_2^2}$, are the critical points at the interior of its domain,

$$\text{int dom } f = \left\{ x \in \mathbb{R}^2 : \frac{x_1^2}{4} + \frac{x_2^2}{16} < 1 \right\},$$

which is just the point 0_2 , together with the points of the boundary of the domain,

$$\text{bd dom } f = \left\{ x \in \mathbb{R}^2 : \frac{x_1^2}{4} + \frac{x_2^2}{16} = 1 \right\}.$$

Concerning $\bar{x} = 0_2$, any $x \in (\text{int dom } f) \setminus \{0_2\}$ satisfies

$$\nabla f(x)^T(x - \bar{x}) = \frac{4x_1^2 + x_2^2}{\sqrt{16 - 4x_1^2 - x_2^2}} > 0,$$

so we have that 0_2 is a strict local minimum of f . Since $\text{dom } f$ is compact, f attains its minimum. The candidates to global minimum are 0_2 and the points of $\text{bd dom } f$, which are obviously dominated by the origin. Consequently, 0_2 is the unique global minimum of f .

In practice, one usually certifies the local optimality of the critical points at the interior of $\text{dom } f$ appealing to either statement (ii) in Theorem 1.25 or to the simplified version of statement (ii) in Theorem 1.27 below consisting in replacing statement (ii) with “the positive definiteness of the Hessian matrix of f at an interior point of $\text{dom } f$ implies that it is a strict local minimum of f ” (direct proofs can be found in many calculus textbooks, e.g., [61, Theorem 28.3.3]. Theorem 1.27 will be used in Chapter 5.

Theorem 1.27 (Second-order optimality conditions) *Let $\bar{x} \in \text{int } F$ be a critical point of f and assume that $f \in C^2(U)$ for some neighborhood U of \bar{x} . The following statements hold:*

- (i) *If \bar{x} is a local minimum of f , then $\nabla^2 f(\bar{x})$ is positive semidefinite.*
- (ii) *If $\nabla^2 f(\bar{x})$ is positive definite, then there exist scalars $\gamma > 0$ and $\varepsilon > 0$ such that*

$$f(x) \geq f(\bar{x}) + \gamma \|x - \bar{x}\|^2, \quad \forall x \in \bar{x} + \varepsilon \mathbb{B}, \quad (1.19)$$

(which implies that \bar{x} is a strict local minimum of f) and, moreover, $\nabla f(x) \neq 0_n$ for all $x \neq \bar{x}$ in certain neighborhood of \bar{x} (which means that \bar{x} is an isolated critical point).

- (iii) *If $\nabla^2 f(\bar{x})$ is indefinite, then \bar{x} is a saddle point.*

Proof Since $f \in C^2(U)$, $\nabla^2 f(\bar{x}) \in S_n$ and its eigenvalues are all real numbers. Let u_1, u_2, \dots, u_n be an orthonormal basis of eigenvectors corresponding to the eigenvalues $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$. From the stationarity condition $\nabla f(\bar{x}) = 0_n$ and (1.10), we get for all $p \in \mathbb{R}^n$ that

$$\begin{aligned} f(\bar{x} + p) &= f(\bar{x}) + \nabla f(\bar{x})^T p + \frac{1}{2} p^T \nabla^2 f(\bar{x}) p + o(\|p\|^2) \\ &= f(\bar{x}) + \frac{1}{2} p^T \nabla^2 f(\bar{x}) p + o(\|p\|^2). \end{aligned} \quad (1.20)$$

Thus, for all $\alpha > 0$, one has by (1.12)

$$f(\bar{x} + \alpha u_1) = f(\bar{x}) + \frac{\lambda_1}{2} \alpha^2 + o(\alpha^2) = f(\bar{x}) + \left(\frac{\lambda_1}{2} + \frac{o(\alpha^2)}{\alpha^2} \right) \alpha^2.$$

- (i) If $\nabla^2 f(\bar{x})$ is not positive semidefinite, one has $\lambda_1 < 0$. Then, there exists $\alpha_0 > 0$ such that

$$\frac{\lambda_1}{2} + \frac{o(\alpha^2)}{\alpha^2} < 0, \quad \forall \alpha \in]0, \alpha_0[.$$

Hence, for all $\alpha \in]0, \alpha_0[$, one has

$$f(\bar{x} + \alpha u_1) < f(\bar{x}), \quad (1.21)$$

implying that \bar{x} is not a local minimum of f .

(ii) The positive definite assumption implies that $\lambda_1 > 0$. Any vector $p \in \mathbb{R}^n$ can be expressed as

$$p = \sum_{i=1}^n \rho_i u_i$$

for certain scalars ρ_1, \dots, ρ_n . Then

$$\nabla^2 f(\bar{x}) p = \sum_{i=1}^n \rho_i \nabla^2 f(\bar{x}) u_i = \sum_{i=1}^n \rho_i \lambda_i u_i,$$

and

$$p^T \nabla^2 f(\bar{x}) p = \left(\sum_{i=1}^n \rho_i u_i^T \right) \left(\sum_{j=1}^n \rho_j \lambda_j u_j \right) = \sum_{i=1}^n \rho_i^2 \lambda_i \|u_i\|^2 = \sum_{i=1}^n \rho_i^2 \lambda_i \geq \lambda_1 \|p\|^2.$$

Combining the latter inequality and (1.20), we get, for all p ,

$$f(\bar{x} + p) - f(\bar{x}) \geq \frac{\lambda_1}{2} \|p\|^2 + o(\|p\|^2) = \left(\frac{\lambda_1}{2} + \frac{o(\|p\|^2)}{\|p\|^2} \right) \|p\|^2.$$

We have thus proved (1.19) for any pair of scalars $\varepsilon > 0$ and $\gamma > 0$ such that

$$\frac{\lambda_1}{2} + \frac{o(\|p\|^2)}{\|p\|^2} \geq \gamma, \quad \text{for all } p \text{ such that } \|p\| \leq \varepsilon.$$

(A possible choice is $\gamma = \frac{\lambda_1}{4}$ together with some $\varepsilon > 0$ such that $\|p\| \leq \varepsilon \Rightarrow \left| \frac{o(\|p\|^2)}{\|p\|^2} \right| < \frac{\lambda_1}{4}$).

Now we show, based on the continuity of $\nabla^2 f$, the existence of an open ball U centered at \bar{x} such that $\nabla^2 f(x)$ is positive definite for all $x \in U$. Indeed, by (1.12) and (1.17), we get

$$\begin{aligned} z^T \nabla^2 f(x) z &= z^T \nabla^2 f(\bar{x}) z + z^T (\nabla^2 f(x) - \nabla^2 f(\bar{x})) z \\ &\geq (\lambda_1 - \|\nabla^2 f(x) - \nabla^2 f(\bar{x})\|) \|z\|^2, \quad \forall z \in \mathbb{R}^n, \end{aligned}$$

By the continuity of $\nabla^2 f$ at \bar{x} and (1.16), there exists an open ball U centered at \bar{x} such that $\|\nabla^2 f(x) - \nabla^2 f(\bar{x})\| < \lambda_1$, for all $x \in U$. Hence, $\nabla^2 f(x)$ is positive definite for all $x \in U$.

Next we prove that $\nabla f(x) \neq 0_n$ for all $x \in U \setminus \{\bar{x}\}$ by contradiction. Assume there exists $\hat{x} \in U \setminus \{\bar{x}\}$ such that $\nabla f(\hat{x}) = 0_n$. Consider the functions $\varphi_i : [0, 1] \rightarrow \mathbb{R}$ such that $\varphi_i(t) := e_i^T \nabla f(\bar{x} + t(\hat{x} - \bar{x}))$, $i = 1, \dots, n$, where e_i denotes the i -th vector of the standard basis. By Barrow's rule (second fundamental theorem of calculus) applied to φ_i at the points 0 and 1, one gets

$$\begin{aligned} e_i^T (\nabla f(\hat{x}) - \nabla f(\bar{x})) &= \varphi_i(1) - \varphi_i(0) = \int_0^1 \varphi_i'(t) dt \\ &= \int_0^1 e_i^T \nabla^2 f(\bar{x} + t(\hat{x} - \bar{x}))(\hat{x} - \bar{x}) dt, \end{aligned}$$

for all $i = 1, \dots, n$, which yields

$$\nabla f(\hat{x}) - \nabla f(\bar{x}) = \int_0^1 \nabla^2 f(\bar{x} + t(\hat{x} - \bar{x}))(\hat{x} - \bar{x}) dt. \quad (1.22)$$

Multiplying both members of (1.22) by $(\hat{x} - \bar{x})^T$, and taking into account that $\nabla f(\bar{x}) = \nabla f(\hat{x}) = 0_n$ and that $\bar{x} + t(\hat{x} - \bar{x}) \in U$ for all $t \in [0, 1]$, we obtain

$$0 = (\hat{x} - \bar{x})^T 0_n = \int_0^1 (\hat{x} - \bar{x})^T \nabla^2 f(\bar{x} + t(\hat{x} - \bar{x}))(\hat{x} - \bar{x}) dt > 0,$$

which is the aimed contradiction.

(iii) If $\nabla^2 f(\bar{x})$ has eigenvalues of different sign, then $\lambda_1 < 0$ and $\lambda_n > 0$. By the same argument as in (i), we deduce that (1.21) holds. A similar argument allows to assert the existence of some $\mu_0 > 0$ such that

$$f(\bar{x} + \mu u_n) = f(\bar{x}) + \left(\frac{\lambda_n}{2} + \frac{o(\mu^2)}{\mu^2} \right) \mu^2 > f(\bar{x}), \quad \forall \mu \in]0, \mu_0[.$$

Hence, \bar{x} is a saddle point. \square

Taking into account the useful information on \bar{x} provided by statement (ii) in Theorem 1.27, its assumptions motivate the following definition.

Definition 1.28 Let f be twice differentiable at $\bar{x} \in \text{int } F$. The point \bar{x} is said to be a *nonsingular local minimum* of f if $\nabla f(\bar{x}) = 0_n$ and $\nabla^2 f(\bar{x})$ is positive definite.

Observe that the second-order sufficient condition does not certify that 0_2 is a local minimum of $f(x) = x_1^4 + x_2^4$ while the first-order sufficient condition does guarantee this. On the contrary, the second-order sufficient condition allows to assert that 0_2 is a strict local minimum of $f(x) = \|x\|^2$ and of the function f of Example 1.10, as $\nabla^2 f(0_2)$ is given by

$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{4} \end{bmatrix}$$

respectively, both matrices being positive definite.

Example 1.29 Consider the unconstrained optimization problem in \mathbb{R}^2

$$P_1 : \text{Min } f(x) = x_1^3 - 12x_1x_2 + 8x_2^3.$$

We have $\nabla f(x) = (3x_1^2 - 12x_2, -12x_1 + 24x_2^2)^T$, so we have that the critical points are $(2, 1)^T$ and $(0, 0)^T$. The Hessian matrix is

$$\nabla^2 f(x) = \begin{bmatrix} 6x_1 & -12 \\ -12 & 48x_2 \end{bmatrix},$$

with

$$\nabla^2 f(2, 1) = \begin{bmatrix} 12 & -12 \\ -12 & 48 \end{bmatrix} \quad \text{and} \quad \nabla^2 f(0, 0) = \begin{bmatrix} 0 & -12 \\ -12 & 0 \end{bmatrix},$$

which are positive definite and not even positive semidefinite, respectively. Thus, the only candidate to be a global minimum of P_1 is $(2, 1)^T$, of which we can only ensure that is a nonsingular local minimum. In fact, $v(P_1) = -\infty$, since $\lim_{x_1 \rightarrow -\infty} f(x_1, 0) = -\infty$.

1.4 Coercivity

Throughout this section, F represents the feasible set of problem P in (1.1). The celebrated Weierstrass theorem establishes that a continuous function $f : F \rightarrow \mathbb{R}$ on a compact set $F \subset \mathbb{R}^n$ attains its minimum on F , that is, there exists an element $\bar{x} \in F$ such that $f(\bar{x}) \leq f(x)$ for all $x \in F$. We will show that the compactness assumption can be relaxed to closedness provided that f satisfies a growth assumption at infinity called “coercivity”.

Definition 1.30 The *sublevel set* λ of f is

$$S_\lambda(f) := \{x \in F : f(x) \leq \lambda\}.$$

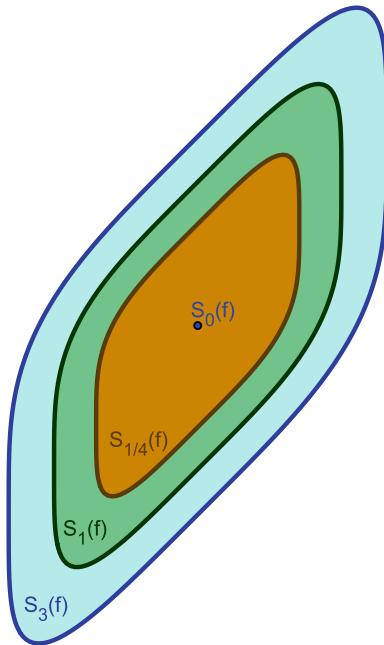
Proposition 1.31 (Attainment of the infimum) *Let $f : F \rightarrow \mathbb{R}$. If there exists a scalar $\lambda \in \mathbb{R}$ such that $S_\lambda(f)$ is a compact set and the restriction of f to the sublevel set $S_\lambda(f)$ is continuous, then f attains its minimum on F .*

Proof Let $\lambda \in \mathbb{R}$ be such that $S_\lambda(f)$ is compact and the restriction of f to $S_\lambda(f)$, represented by $f|_{S_\lambda(f)}$, is continuous. The points of $F \setminus S_\lambda(f)$ are dominated by the points of $S_\lambda(f)$, so it is enough to show that $f|_{S_\lambda(f)}$ attains its minimum on $S_\lambda(f)$, but this is a straightforward consequence of the Weierstrass theorem. \square

Corollary 1.32 *Let P be the optimization problem in (1.1), with $\emptyset \neq C \subset \mathbb{R}^n$ closed and $\{f; h_i, i = 1, \dots, m; g_j, j = 1, \dots, p\} \subset \mathcal{C}(C)$. If there exists a scalar $\lambda \in \mathbb{R}$ such that $S_\lambda(f)$ is a nonempty bounded set, then P is solvable.*

Proof The same argument as in the second part of the proof of Proposition 1.2 shows that $S_\lambda(f)$ is closed. The conclusion comes from Proposition 1.31. \square

Fig. 1.15 Sublevel sets of the function
 $f(x) = (x_1 - x_2)^4 + (x_1 - 1)^4$



Example 1.33 Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be the function $f(x) = (x_1 - x_2)^4 + (x_1 - 1)^4$. Since f is continuous, $S_\lambda(f)$ is closed for all $\lambda \in \mathbb{R}$. Next we show that the sublevel sets are also bounded. Let $\lambda \in \mathbb{R}$ be such that $S_\lambda(f) \neq \emptyset$ (which implies $\lambda > 0$). Since $f(x) \leq \lambda$, $(x_1 - x_2)^4 \leq \lambda$ and $(x_1 - 1)^4 \leq \lambda$, we have $|x_1 - x_2| \leq \sqrt[4]{\lambda}$ and $|x_1 - 1| \leq \sqrt[4]{\lambda}$. Thus, $|x_1| \leq 1 + \sqrt[4]{\lambda}$ and $|x_2| \leq 1 + 2\sqrt[4]{\lambda}$ for all $x \in S_\lambda(f)$. In fact,

$$S_\lambda(f) \subset \left\{ x \in \mathbb{R}^2 : -\sqrt[4]{\lambda} \leq x_1 - x_2 \leq \sqrt[4]{\lambda}, -\sqrt[4]{\lambda} \leq x_1 - 1 \leq \sqrt[4]{\lambda} \right\},$$

which is a bounded set. The boundedness of $S_\lambda(f)$ for any $\lambda \geq 0$ guarantees the existence of a global minimum that is obviously the unique zero of f , which is the point $(1, 1)^T$ (see Fig. 1.15).

1.4.1 Coercive Functions

Definition 1.34 Let $F \subset \mathbb{R}^n$ be unbounded. The function $f : F \rightarrow \mathbb{R}$ is *coercive* on F whenever

$$\lim_{\|x\| \rightarrow +\infty, x \in F} f(x) = +\infty.$$

Example 1.35 Given $u \in \mathbb{R}^n$ and $m \in \mathbb{N}$, the function $f(x) = \|x - u\|^m$ is coercive on \mathbb{R}^n . Indeed, due to $\|x\| \leq \|x - u\| + \|u\|$ (by the triangular inequality), one has

$$\lim_{\|x\| \rightarrow +\infty} f(x) \geq \lim_{\|x\| \rightarrow +\infty} (\|x\| - \|u\|)^m = +\infty.$$

Obviously, the unique global minimum of f is u .

Example 1.36 The function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ such that $f(x) = x_1^2 - 2x_1x_2 + x_2^2$ is not coercive because $f(x_1, x_1) = 0$ for all $x_1 \in \mathbb{R}$, although $\lim_{|x_1| \rightarrow +\infty} f(x_1, x_2) = +\infty$ for all $x_2 \in \mathbb{R}$ and $\lim_{|x_2| \rightarrow +\infty} f(x_1, x_2) = +\infty$ for all $x_1 \in \mathbb{R}$. As $f(x) = (x_1 - x_2)^2$, the set of global minima of f is the line $x_2 = x_1$.

It is easy to give examples of continuous functions on \mathbb{R}^n that are not coercive and have some bounded sublevel sets (e.g., if $f(x) = 1 - \exp(-x^2)$, then $S_\lambda(f)$ is bounded for all $\lambda < 1$ while $S_\lambda(f) = \mathbb{R}$ for $\lambda \geq 1$).

Proposition 1.37 (Characterizing coercivity) *Let $F \subset \mathbb{R}^n$ be unbounded. Then, f is coercive on F if and only if $S_\lambda(f)$ is bounded for all $\lambda \in \mathbb{R}$.*

Proof We shall prove the equivalence of the negations.

Assume that $S_\lambda(f)$ is unbounded for some $\lambda \in \mathbb{R}$. Let $\{x_r\} \subset S_\lambda(f)$ be such that $\lim_{r \rightarrow \infty} \|x_r\| = +\infty$. Since $f(x_r) \leq \lambda$ for all $r \in \mathbb{N}$, f cannot be coercive.

Conversely, assume that f is not coercive. Then we can take some $k > 0$ such that, for all $r \in \mathbb{N}$, there exists a point x_r such that $\|x_r\| \geq r$ and $f(x_r) \leq k$. However, $\{x_r\} \subset S_k(f)$ implies the unboundedness of $S_k(f)$. \square

Theorem 1.38 (Existence of global minimum) *If $\emptyset \neq C \subset \mathbb{R}^n$ is closed,*

$$\{f; h_i, i = 1, \dots, m; g_j, j = 1, \dots, p\} \subset \mathcal{C}(C)$$

and either F is bounded or f is coercive on F , then P has some global minimum.

Proof Assume that F is unbounded. Taking an arbitrary feasible solution $x_1 \in F$, we have $S_{f(x_1)}(f) \neq \emptyset$. The conclusion follows from Corollary 1.32 and Proposition 1.37. \square

Example 1.39 The objective function of problem P_3 in Example 1.6 is $f(x) = e^{x_1+x_2} + e^{-x_1} + e^{-x_2}$. Obviously, for $\lambda \geq 1$ such that $S_\lambda(f) = \emptyset$, $x \in S_\lambda(f)$ implies that $e^{x_1+x_2} \leq \lambda$, $e^{-x_1} \leq \lambda$ and $e^{-x_2} \leq \lambda$, i.e., $x_1 + x_2 \leq \ln \lambda$, $-x_1 \leq \ln \lambda$ and $-x_2 \leq \ln \lambda$. Since

$$S_\lambda(f) \subset (\ln \lambda)C,$$

where C is the region limited by the triangle of vertices $(-1, -1)^T$, $(2, -1)^T$, and $(-1, 2)^T$, that is,

$$C = \left\{ \mu_1 \begin{pmatrix} -1 \\ -1 \end{pmatrix} + \mu_2 \begin{pmatrix} 2 \\ -1 \end{pmatrix} + \mu_3 \begin{pmatrix} -1 \\ 2 \end{pmatrix} : \mu_1, \mu_2, \mu_3 \geq 0, \sum_{i=1}^3 \mu_i = 1 \right\},$$

one has that $S_\lambda(f)$ is bounded for all λ , f is coercive and P_3 has some global minimum.

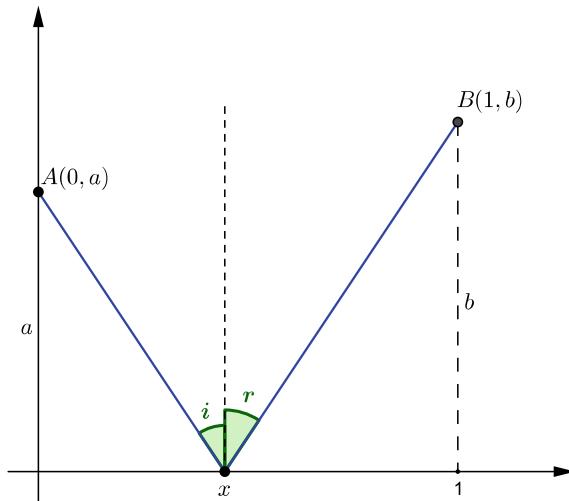
The next example illustrates the use of optimization in natural sciences (in particular, in physics).

Example 1.40 According to Pierre de Fermat, the path taken by a ray of light between two points A and B is the path that can be traversed in the least time. We will use this principle to derive the law of specular reflection, which states that the incident ray, the reflected ray, and the normal line to the mirror at the incidence point, are coplanar and both rays make the same angle with respect to the normal line. In fact, by a simple geometrical argument, the light will describe a path formed by a first segment linking A with some (incidence) point of the mirror and a second segment linking the incidence point with B . Moreover, these two segments will be contained in the plane M orthogonal to the mirror which contains A and B . The first step consists of selecting a suitable 2D geometric model allowing to compute, in a second step, the incidence point by solving an optimization problem. We take as x -axis the intersection of M with the mirror and as y -axis the orthogonal line to the mirror from A . Moreover, we take as unit of length the distance between the projections of A and B onto the mirror. Hence, we can assume that the incoming light travels from $A(0, a)$ to an unknown point $(x, 0)$, while the outgoing light arrives to $B(1, b)$, where $a > 0$ and $b > 0$ (see Fig. 1.16). As the light travels through an homogeneous medium with constant velocity, the optimization problem consists of minimizing the path length, i.e.,

$$P_1 : \text{Min } f(x) = \sqrt{x^2 + a^2} + \sqrt{(x - 1)^2 + b^2}.$$

Since f is continuous and coercive on \mathbb{R} , the existing optimal solution must satisfy the first-order necessary condition

Fig. 1.16 Law of reflection of light



$$f'(x) = \frac{x}{\sqrt{a^2 + x^2}} + \frac{x - 1}{\sqrt{(x - 1)^2 + b^2}} = 0.$$

Eliminating the square roots (operation which could introduce false critical points) and simplifying, one gets

$$b^2 x^2 = a^2 (x - 1)^2.$$

Taking square roots of both sides of this equation, we obtain

$$x = \frac{a}{a + b} \quad \text{and} \quad x = \frac{a}{a - b}.$$

It is easy to see that the unique solution of $f'(x) = 0$ is

$$\bar{x} = \frac{a}{a + b},$$

which is then the unique optimal solution to P_1 . Denoting by i and r the incidence and reflection angles, we have

$$\tan i = \frac{1}{a + b} = \tan r.$$

Due to the injectivity of the tangent function on $]-\frac{\pi}{2}, \frac{\pi}{2}[$, we thus get the law of specular reflection previously stated by Descartes in his *Dioptrique* (and before him by the Greek mathematician and engineer Heron of Alexandria) on empirical basis: the angle of incidence equals the angle of reflection.

Refraction is the change in direction of propagation of light due to a change in its transmission medium separated by a plane interface. The problem to be solved is now

$$P_2 : \text{Min } f(x) = \frac{\sqrt{x^2 + a^2}}{v_1} + \frac{\sqrt{(x - 1)^2 + b^2}}{v_2},$$

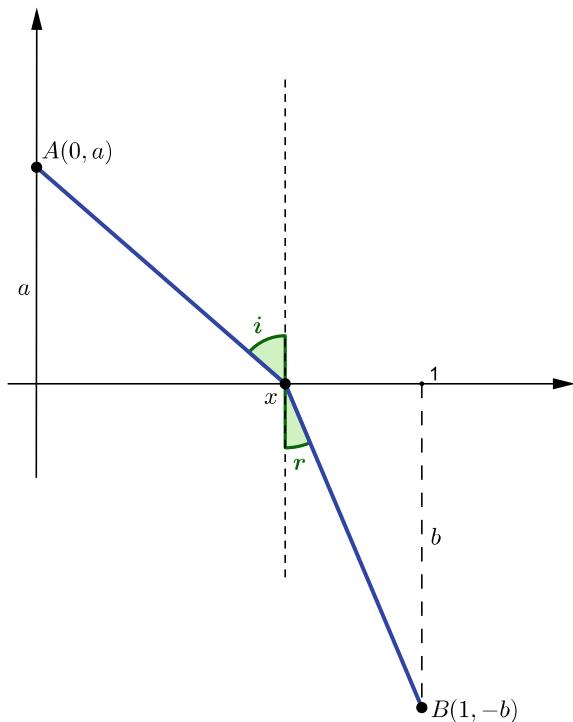
where v_1 and v_2 represent the velocity of the light at medium 1 and medium 2, respectively (see Fig. 1.17). A similar argument to that of the proof of the reflection law provides the equation of degree four to be satisfied by the critical points. Fortunately, we do not need to solve such a difficult equation. In fact, it is enough to observe that the critical point \bar{x} determines incidence and refraction angles, i and r , whose respective sinuses are $\sin i = \frac{\bar{x}}{\sqrt{a^2 + \bar{x}^2}}$ and $\sin r = \frac{1 - \bar{x}}{\sqrt{(\bar{x} - 1)^2 + b^2}}$, so we have that

$$\frac{\sin i}{v_1} = \frac{\sin r}{v_2}.$$

This law was mistakenly stated by Descartes (following the experimental physicist Snell) as $\frac{i}{v_1} = \frac{r}{v_2}$.

It is worth observing that the Fermat principle should be reformulated more precisely in this way: the path taken between two points A and B by a ray of light is

Fig. 1.17 Law of refraction of light



the path that provides a critical minimum for the time necessary to travel from A to B . Fermat introduced the concept of derivative to solve optimization problems and can be considered as one of the founders of differential calculus together with his contemporaries Newton and Leibnitz.

Example 1.41 A fishing Western company dispatches a warship to protect its float at the Indian Ocean from pirate-hunting. This float is formed by m fish boats currently placed at x_1, \dots, x_m . The optimization problem to be solved consists of locating the warship in such a way that the response time for a possible request of protection from the furthest fish boat is minimized. Denoting by x the position of the warship, the problem can be formulated as

$$P_1 : \text{Min}_{x \in \mathbb{R}^2} f_1(x) := \max\{\|x - x_i\| : i = 1, \dots, m\},$$

whose objective function f_1 is continuous and coercive on \mathbb{R}^2 , so P_1 has some global minimum. However, it is not easy to compute candidates to global minimum as f_1 is not differentiable. Recalling that the maximum of a finite set of numbers is the least upper bound, P_1 can be reformulated as minimizing the common upper bound t for $\|x - x_i\|$, $i = 1, \dots, m$, i.e.,

$$P_2 : \text{Min}_{(x,t) \in \mathbb{R}^3} f_2(x, t) := t \\ \text{s.t.} \quad t \geq \|x - x_i\|, i = 1, \dots, m.$$

Unfortunately, this reformulation has the same drawback as P_1 , as the constraint functions are still nondifferentiable. This inconvenience can be removed by taking squares, as the equivalent problem

$$\begin{aligned} P_3 : \text{Min}_{(x,t) \in \mathbb{R}^3} \quad & f_3(x, t) := t \\ \text{s.t.} \quad & t^2 \geq \|x - x_i\|^2, \quad i = 1, \dots, m, \end{aligned}$$

has linear objective function and quadratic constraints. However, the constraint functions are not convex (the reader is referred to the convexity test for \mathcal{C}^2 -functions in the next chapter), so the local minima computed by numerical methods are not necessarily global minima. This inconvenience also disappears after two new reformulations. Taking the square of the objective function, one gets

$$\begin{aligned} P_4 : \text{Min}_{(x,t) \in \mathbb{R}^3} \quad & f_4(x, t) := t^2 \\ \text{s.t.} \quad & t^2 \geq \|x - x_i\|^2, \quad i = 1, \dots, m, \end{aligned}$$

and replacing the decision variable t by $y = t^2$, we get our favorite reformulation

$$\begin{aligned} P_5 : \text{Min}_{(x,y) \in \mathbb{R}^3} \quad & f_5(x, y) := y \\ \text{s.t.} \quad & y \geq \|x - x_i\|^2, \quad i = 1, \dots, m, \end{aligned}$$

which has again a linear objective function and quadratic constraints, but the latter are all convex. The moral of this example is that different optimization models can be proposed for a given real problem, each one with its corresponding advantages and disadvantages, e.g., P_1 allows to easily show the existence of a global minimum while P_5 allows its computation in the easiest way.

Example 1.42 Let $u, v \in \mathbb{R}^n$ be two nonzero vectors. Consider the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}_+$ such that $f(x) = \|x_1 u + x_2 v\|$. Obviously, the set of global minima of f is $F^* = \{x \in \mathbb{R}^2 : x_1 u + x_2 v = 0_n\}$, so there exists a unique global minimum if and only if $\{u, v\}$ is linearly independent. Next we show that f is coercive if and only if $\{u, v\}$ is linearly independent.

(a) Suppose that $\{u, v\}$ is linearly dependent. Let $\alpha, \beta \in \mathbb{R}$ be not simultaneously null, so $\alpha u + \beta v = 0_n$. Then, given $r \in \mathbb{N}$, one has $f(\alpha r, \beta r) = \|r(\alpha u + \beta v)\| = 0$, null, with $\|(\alpha r, \beta r)^T\| = r\sqrt{\alpha^2 + \beta^2}$. Therefore, we cannot have $\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty$, i.e., f is not coercive.

(b) Suppose now that f is not coercive. Then there exists $k \in \mathbb{R}$ such that the sublevel set $S_k(f)$ is unbounded. Take a sequence $\{(\alpha_r, \beta_r)^T\} \subset S_k(f)$ such that $\|(\alpha_r, \beta_r)^T\| \rightarrow +\infty$. Since $\|\alpha_r u + \beta_r v\| = f(\alpha_r, \beta_r) \leq k$ for all r , the sequence $\{\alpha_r u + \beta_r v\}$ is bounded. Dividing the elements of $\{\alpha_r u + \beta_r v\}$ by $\|(\alpha_r, \beta_r)^T\|$ and taking $\lim_{r \rightarrow \infty}$, we get

$$\lim_{r \rightarrow \infty} \left(\frac{\alpha_r}{\|(\alpha_r, \beta_r)^T\|} u + \frac{\beta_r}{\|(\alpha_r, \beta_r)^T\|} v \right) = 0_n. \quad (1.23)$$

Since the sequence $\left\{ \left(\frac{\alpha_r}{\|(\alpha_r, \beta_r)^T\|}, \frac{\beta_r}{\|(\alpha_r, \beta_r)^T\|} \right)^T \right\}$ is contained in the unit circle, which is a compact set, we can assume without loss of generality the existence of a point $(\alpha, \beta)^T$ in the unit circle such that

$$\lim_{r \rightarrow \infty} \left(\frac{\alpha_r}{\|(\alpha_r, \beta_r)^T\|}, \frac{\beta_r}{\|(\alpha_r, \beta_r)^T\|} \right)^T = (\alpha, \beta)^T. \quad (1.24)$$

Combining (1.23) and (1.24) we conclude that $\alpha u + \beta v = 0_n$, with $\alpha^2 + \beta^2 = 1$, so we have that $\{u, v\}$ is linearly dependent.

1.4.2 The Fundamental Theorem of Algebra*

We start by recalling some basic facts concerning complex numbers and complex polynomials. The field of complex numbers is $\mathbb{C} = \mathbb{R}^2$ equipped with the componentwise sum and a product which is better described by writing any complex number $x = (x_1, x_2)$ in the binary form $x = x_1 + ix_2$, where i denotes the square root of -1 , i.e., $i^2 = -1$. In this way, given $x, z \in \mathbb{C}$,

$$\begin{aligned} xz &= (x_1 + ix_2)(z_1 + iz_2) = (x_1z_1 - x_2z_2) + i(x_1z_2 + x_2z_1) \\ &= (x_1z_1 - x_2z_2, x_1z_2 + x_2z_1). \end{aligned}$$

Moreover, the Euclidean norm of $x \in \mathbb{C}$, $\|x\|$, is here called absolute value, and it is denoted by $|x|$. Thus, the absolute value satisfies the well-known properties of the norm together with $|xz| = |x||z|$ for all $x, z \in \mathbb{C}$. It is also easy to obtain the roots of any index of a given complex number $x \neq 0$ from its trigonometric form $x = |x|e^{i\theta}$, where $e^{i\theta} = \cos \theta + i \sin \theta$, as $x^m = |x|^m e^{im\theta}$ for any $m \in \mathbb{N}$. From the topological perspective, \mathbb{C} is equipped with the topology derived from the Euclidean norm, so that the limits of complex functions can be computed by the same rules as the limits of functions from \mathbb{R}^2 to \mathbb{R}^2 together with a new one for the quotient (operation that does not exist for functions from \mathbb{R}^n to \mathbb{R}^n when $n \geq 3$). For instance, if $f, g : \mathbb{C} \rightarrow \mathbb{C}$ satisfy $\lim_{z \rightarrow a} f(z) = c$ and $\lim_{z \rightarrow a} g(z) = d$, then $\lim_{z \rightarrow a} (fg)(z) = cd$ and $\lim_{z \rightarrow a} \left(\frac{f}{g} \right)(z) = \frac{c}{d}$ provided that $g(z) \neq 0$ for $z \neq a$ and $d \neq 0$.

Computing the zeros of a function $f : \mathbb{C} \rightarrow \mathbb{C}$ can be reduced to the problem of computing the global minima of $|f| : \mathbb{C} \rightarrow \mathbb{R}$, whose graph, $\text{gph}|f|$, admits a geometrical representation as it is contained in \mathbb{R}^3 (on the contrary, we cannot represent $\text{gph } f \subset \mathbb{R}^4$). If f is continuous, $|f|$ is also continuous, but the differentiability of f (as a function of two variables) is not inherited by $|f|$. This inconvenience (from the computational point of view) can be avoided by replacing the objective function $|f|$ with its square $|f|^2$.

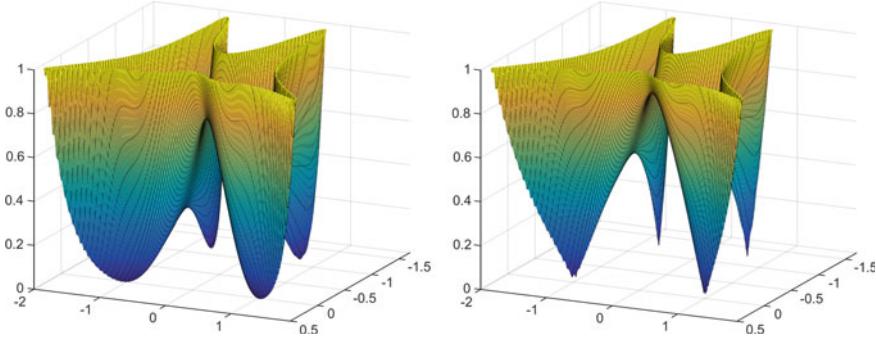


Fig. 1.18 Graphs of $|f|^2$ (left) and of $|f|$ (right)

Example 1.43 Consider the function $f(z) = z^4 + z^3 + 2z^2 + z + 1$. Writing $z = x + iy$, $|f(z)|^2$ can be expressed as $(1 + x + 2x^2 - 2y^2 + x^3 - 3xy^2 + x^4 - 6x^2y^2 + y^4)^2 + (y + 4xy + 3x^2y - y^3 + 4x^3y - 4xy^3)^2$. Figure 1.18 shows portions of $\text{gph}|f|$ and $\text{gph}|f|^2$. Observe that both functions are continuous, but only $|f|^2$ is differentiable. The global minima of $|f|^2$ (and $|f|$) are $\pm i$ and $\frac{-1 \pm i\sqrt{3}}{2}$ (points where $\text{gph}|f|^2$ is tangent to the plane $\mathbb{R}^2 \times \{0\}$ while $\text{gph}|f|$ is sharp at these points), which are the zeros of f as the optimal value of the optimization problem is 0.

The fundamental theorem of algebra, stated by d'Alembert in 1746 and correctly proved by Gauss in 1799, establishes that every nonconstant single-variable polynomial with complex coefficients has at least one complex root. Consequently, by the division algorithm, if $P(z)$ is a complex polynomial of degree $n \geq 1$, it has exactly n zeros on \mathbb{C} (possibly repeated). The basic ingredient of the next proof (proposed in [86]) is the existence of global minima for continuous coercive functions (other simple proofs of the fundamental theorem of algebra are commented, e.g., in [72]).

Theorem 1.44 (Fundamental theorem of algebra) *Any algebraic equation of degree $n \geq 1$ with complex coefficients has some solution in \mathbb{C} .*

Proof The result is trivial for the following four types of equations $P(z) = a_0 + a_1z + \dots + a_nz^n = 0$, with $a_n \neq 0$:

I. When $n = 1$.

II. When $n = 2$.

III. When $n \geq 3$ and $a_0 = 0$, as $z = 0$ is a trivial solution.

IV. When $n \geq 3$ and $a_1 = \dots = a_{n-1} = 0$, as we know how to solve the equation $P(z) = a_0 + a_nz^n = 0$ (whose solutions are the roots of index n of $-\frac{a_0}{a_n}$).

Thus, we will assume that the given polynomial can be expressed as

$$P(z) = a_0 + a_kz^k + \dots + a_nz^n, \quad 1 \leq k < n,$$

with $n \geq 3$ and $a_0 \neq 0 \neq a_k$.

We first prove the coercivity, on \mathbb{R}^2 , of the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}_+$ such that $f(z) := |P(z)|$ for all $z \in \mathbb{R}^2$, i.e., that $\lim_{|z| \rightarrow +\infty} f(z) = +\infty$. By the well-known properties of the absolute value, we have

$$\begin{aligned} f(z) &= |a_n z^n - (-a_k z^k - \dots - a_0)| \geq |a_n z^n| - |a_k z^k + \dots + a_0| \\ &\geq |a_n| |z|^n - |a_k| |z|^k - \dots - |a_0|, \end{aligned}$$

which together with

$$\lim_{x \rightarrow \infty} (|a_n| x^n - |a_k| x^k - \dots - |a_0|) = +\infty,$$

shows that $\lim_{|z| \rightarrow +\infty} f(z) = +\infty$. Thus, f is coercive.

Obviously, $\text{gph } f$ lies above $\mathbb{R}^2 \times \{0\}$ (the horizontal plane through the origin). We next show that the contact points of $\text{gph } f$ with such a plane are the zeros of P whose existence we are trying to prove.

Let \widehat{z} be a global minimum of f on \mathbb{R}^2 . Then 0 is also a global minimum of $h(z) := f(z + \widehat{z})$ on \mathbb{R}^2 . Obviously, $h(z) = |P(z + \widehat{z})|$ and we can write $P(z + \widehat{z}) = b_0 + b_k z^k + \dots + b_n z^n$, with $1 \leq k < n$ and $b_n = a_n \neq 0$ (this k could not coincide with the previous one). It remains to prove that $b_0 = 0$ as, in this case, since $h(0) = 0$, we have that 0 will be a solution of $P(z + \widehat{z}) = 0$, meaning that \widehat{z} is a solution of $P(z) = 0$.

Since $b_0 + b_k z^k = 0$ has solution in \mathbb{C} , there exists $u \in \mathbb{C}$ such that

$$b_0 + b_k u^k = 0. \quad (1.25)$$

From (1.25) and the triangular property of the absolute value, given $t \in]0, 1[$, one has

$$\begin{aligned} |b_0| &= h(0) \leq h(tu) = |b_0 + b_k t^k u^k + \dots + b_n t^n u^n| \\ &= |b_0 - b_0 t^k + t^k (b_{k+1} t u^{k+1} + \dots + b_n t^{n-k} u^n)| \\ &= |b_0(1 - t^k) + t^k \theta(t)| \leq |b_0(1 - t^k)| + |t^k \theta(t)| \\ &= |b_0|(1 - t^k) + t^k |\theta(t)|, \end{aligned} \quad (1.26)$$

where $\theta(t) := t(b_{k+1} u^{k+1} + \dots + b_n t^{n-k-1} u^n)$. Obviously, $\lim_{t \searrow 0} \theta(t) = 0$.

From (1.26) one gets

$$t^k |b_0| \leq t^k |\theta(t)|,$$

which implies (as $t > 0$)

$$|b_0| \leq |\theta(t)|.$$

Thus, taking $\lim_{t \searrow 0}$ we get the conclusion that $|b_0| = 0$, i.e., $b_0 = 0$. \square

1.5 Exercises

1.1 A civil engineer needs to build a bridge connecting the two sides of a river. Once an appropriate coordinate reference is fixed, he observes that, in the narrowest part of the river, both sides are approximately described by the equations $y = x - 5$ and $y = x^2$. Where must the bridge be constructed so that its length is as small as possible?

1.2 The government of a certain country decides to build a road connecting the cities A and B. City B is located 30 km east and 50 km north from A. A mountain range, which has an approximately constant width of 10 km, runs from north to south equidistant between the two cities. The unit cost of the construction on the mountain is $\sqrt{1.6}$ times higher than on the plane. What is the most economical layout?

1.3 A manufacturing company transforms raw material to produce two classes of products. Each unit of processed raw material provides 2 units of product A and 1 unit of product B, not being required to sell the entire production. If x_1 and x_2 are, respectively, the units of A and B that are put on sale, the selling price will be $40 - x_1$ dollars per unit of A and $30 - x_2$ dollars per unit of B. The unit cost of purchasing and processing raw materials is \$4. Plan the activities of the company.

1.4 A farmer needs to transport 400 m^3 of cereal from his barn to the large silo of a cooperative using his pickup truck on which he is planning to screw an open-top box. He has to order the box to his blacksmith friend, who will only charge him the sheet metal used to build it. Each journey to the silo costs 10 c.u. (currency units) and the price of the sheet metal is 100 c.u./m^2 for the bottom of the box and the double amount for the sides. How should he design the box?

1.5 How a 1 liter cylindrical container should be designed in order to minimize the amount of material used in its construction?

1.6 An object is removed from the refrigerator at 5°C and left in an atmosphere at 25°C . Then, the temperature of the object is observed from time to time:

$t(\text{min})$	0	2	4	8	12	16
$T(\text{ }^\circ\text{C})$	5	11	16	20	21	24

Knowing that, according to Newton's Law of Cooling, the rate of change in the temperature of an object is proportional to the difference between its own temperature and the ambient temperature:

(a) Find a theoretical model (dependent on certain parameters) that describes how the temperature varies with time.

Hint: To solve the differential equation $\frac{dy}{dx} = \frac{f(x)}{g(y)}$, simply separate the variables (by $f(x)dx = g(y)dy$) and integrate both members.

(b) Adjust the theoretical model to the observed data (that is, estimate the parameters). For that, make a change of variables that allow you to minimize, in some way, the

vector of errors (differences between observed and predicted values), arguing the appropriateness of different norms.

Hint: It can be convenient to make a change of variables that simplifies calculation.

1.7 A total of m warehouses must be placed in the most convenient way to supply n stores of a commercial network established in a city whose streets are parallel or orthogonal to each other (like Manhattan). To simplify, we suppose that the stores sell a single product. The store j is located in (a_j, b_j) (coordinates with respect to a reference system whose axes are parallel to the street layout) and its daily demand is r_j units. With respect to the warehouses, the capacity of the warehouse i is c_i units, and we must decide its location, (x_i, y_i) , as well as the number of daily units to be transported during retail hours from the warehouse i to the store j , w_{ij} . Every night the warehouses will be restocked. Construct a model that allows to minimize the supply costs for the stores, assuming that the unit transport costs are proportional to the distances traveled from the warehouse to the corresponding store. Classify the model and say if you can ensure a priori that this model is solvable.

1.8 Let x_1, \dots, x_n be decision variables in a mathematical optimization problem. Prove the equivalence between the following two constraint sets:

- (a) $x_i \in \{0, 1\}, i = 1, \dots, n; x_i + x_j \leq 1, i \neq j, i, j = 1, \dots, n.$
- (b) $x_i^2 - x_i = 0, i = 1, \dots, n; x_i x_j = 0, i \neq j, i, j = 1, \dots, n.$

Compare the advantages and disadvantages of both formulations (as a consequence of the number and type of constraints).

1.9 Study if the constraint $z \in \{0, 1\}$ can be replaced, when it appears in an optimization problem, by

$$z + w = 1, z \geq 0, w \geq 0, zw = 0, z \in \mathbb{R}, w \in \mathbb{R},$$

where w is an auxiliary variable. Compare the advantages and disadvantages of both formulations, whether they are equivalent or not.

1.10 Perform a variable transformation that allows to formulate the following problem as an unconstrained optimization problem:

$$\begin{aligned} & \text{Min } f(x) \\ \text{s.t. } & \sum_{i=1}^n x_i^2 = 1. \end{aligned}$$

Analyze the advantages and disadvantages of the new formulation.

1.11 Suppose that you try to solve the following problem with a software that does not admit inverse trigonometric functions:

$$\begin{aligned} & \text{Min } \arctan(x^2 - 1) + \frac{1}{x} \\ \text{s.t. } & 1 \leq x \leq 2. \end{aligned}$$

Reformulate the problem so that it can be solved.

1.12 Reformulate the following nonlinear optimization problem as a linear one:

$$\begin{aligned} \text{Min } & \exp(2x) / \exp(3y + u^2) \\ \text{s.t. } & x - 2y + 3u^2 \leq 5, \\ & \tan(xyz^2) = 1, \\ & \ln(2x + y) \geq 0, \\ & x \geq 0, y \geq 0. \end{aligned}$$

1.13 A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is *separable* when it can be written as $f(x) = \sum_{i=1}^n f_i(x_i)$, with $f_i : \mathbb{R} \rightarrow \mathbb{R}$ for all $i = 1, 2, \dots, n$. An optimization problem is *separable* when all the functions involved are separable. Transform the following model into a separable one:

$$\begin{aligned} \text{Min } & 20x_1 + 16x_2 - 2x_1^2 - x_2^2 - (x_1 + x_2)^2 \\ \text{s.t. } & x_1 x_2 \leq 5, \\ & x_1 \geq 1, x_2 \geq 2. \end{aligned}$$

1.14 Let $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ be differentiable functions at a point $a \in \mathbb{R}^n$ where $f(a) = g(a)$ and $f(x) \leq g(x)$ for all x in some neighborhood of a . Prove that $\nabla f(a) = \nabla g(a)$.

1.15 Find the local and the global minima of the functions

- (a) $f(x, y) = x^2 + 2x + 3y^2 + 2xy + 10y + 9$;
- (b) $f(x, y) = (x - y)^4 + (x - 1)^4 + 25$.

To this aim, use the following.

- (i) The first-order optimality conditions (exclusively).
- (ii) The first-order necessary condition combined with the second-order sufficient one.
- (iii) The first-order necessary condition combined with coercivity arguments.

1.16 Classify the origin as either local (global) optimum or saddle point of the following functions on \mathbb{R}^2 :

- (a) $f(x_1, x_2) = x_1^2 + x_2^3$.
- (b) $f(x_1, x_2) = x_1^2 + x_2^4$.
- (c) $f(x_1, x_2) = x_1^5 - x_1 x_2^6$.

1.17 Let $f(x, y, z) = e^x + e^y + e^z + 2e^{-x-y-z}$. Prove that:

- (a) $\nabla^2 f$ is positive definite on \mathbb{R}^3 ;
- (b) $(\frac{1}{4} \ln 2, \frac{1}{4} \ln 2, \frac{1}{4} \ln 2)^T$ is the only global minimum of f on \mathbb{R}^3 .

1.18 Find the local and the global minima of the following functions on \mathbb{R}^2 .

- (a) $f(x) = x^T \begin{bmatrix} 1 & -8 \\ 1 & 1 \end{bmatrix} x$.
- (b) $f(x_1, x_2) = x_1^4 + x_2^4 - x_1^2 - x_2^2 + 1$.
- (c) $f(x_1, x_2) = 12x_1^3 - 36x_1x_2 - 2x_2^3 + 9x_2^2 - 72x_1 + 60x_2$.
- (d) $f(x_1, x_2) = x_1^2 - 4x_1 + 2x_2^2$.

- (e) $f(x_1, x_2) = e^{-(x_1^2 + x_2^2)}$.
 (f) $f(x_1, x_2) = x_1^3 + e^{3x_2} - 3x_1 e^{x_2}$.
 (g) $f(x_1, x_2) = (x_1^2 x_2 - x_1 - 1)^2 + (x_1^2 - 1)^2$.
 (h) $f(x_1, x_2) = (x_1 x_2 - 1)^2 + x_1^2$.
 (i) $f(x_1, x_2) = x_1^2 - \sin x_2$.

1.19 Find the maxima and the minima of the following functions on \mathbb{R}^3 .

- (a) $f(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 - 4x_1 x_2$.
 (b) $f(x_1, x_2, x_3) = (2x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - 1)^2$.

1.20 Find the local and the global minima of the function $f(x_1, x_2) = x_1^3 - 3ax_1 x_2 + x_2^3$ for all the possible values of $a \in \mathbb{R}$.

1.21 (a) Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be continuous and let $\bar{x} \in \mathbb{R}$. Prove that the following statements are equivalent:

- (i) \bar{x} is a global minimum of f .
 (ii) $f(x) = f(\bar{x}) \implies x$ is a local minimum of f .
 (b) Is the continuity of $f : \mathbb{R} \rightarrow \mathbb{R}$ essential in (a)?
 (c) Does the equivalence (i) \iff (ii) holds for $f : \mathbb{R}^n \rightarrow \mathbb{R}$?

1.22 In order to save the world from destruction, Agent 007 must disable a nuclear bomb placed on an islet located 50 m away from the beach on whose rectilinear shore he is resting, at a distance of 100 meters from the point of the beach closest to the islet. The agent has 74 s to reach the bomb and disable it, operation that requires 30 s. His running speed is 5 m/s whereas his swimming speed is 2 m/s. Should he try to carry out the deactivation or wait for the apocalypse while he drains the beer he is drinking in good company?

1.23 The City Council of a tourist town is considering installing several bars on a pristine beach, which is perfectly straight and extremely narrow and has a length of 1 km. It is known that bathers are distributed on the shore beach at random (with a uniform distribution). Bathers are supposed to have their drinks at the nearest bar and the daily number of visits to the bar does not depend on the distance.

Suppose that the City Council authorizes a single bar to the best bidder.

(a) Where should it be installed to minimize the total distance walked around by bathers? What would the average distance walked around by bathers be?

Suppose now that the City Council decides to install a second bar after the first one was installed in the place determined in (a).

(b) If the City Council decides to install them at the ends of the beach, how many km will bathers walk around on average to quench their thirst?

(c) If the City Council, with philanthropic mood, wants to minimize the total distance walked around by all the bathers, where should the bars be located? How much effort would this decision save compared to that of (b)?

Part I

Analytical Optimization

Chapter 2

Convexity



This chapter starts with the study of basic properties of convex sets, including the separation and the supporting hyperplane theorems. Next, we introduce a class of functions that are very useful in optimization, the convex functions, for which the local minima are also global minima. We begin with functions of one variable and follow with functions of several variables (or multivariate). The chapter finishes with the study of two particular classes of convex functions, the strictly convex and the strongly convex functions, which are introduced in order to ensure the existence and uniqueness of global minimum. Taking into account the instrumental character of this chapter, we have preferred to sacrifice generality to simplicity in the presentation of some results; compare, e.g., Propositions 2.33, 2.37, and 2.60 below with [80, Theorem 23.4], [80, Theorem 24.7], and [25, Theorem 2.8], respectively.

2.1 Convex Sets

Definition 2.1 A set $C \subset \mathbb{R}^n$ is *convex* when $(1 - \mu)x + \mu y \in C$ for all $x, y \in C$ and all $\mu \in]0, 1[$.

In particular, \emptyset and \mathbb{R}^n are convex sets.

Definition 2.2 The *convex hull* of a set $X \subset \mathbb{R}^n$, denoted by $\text{conv } X$, is defined as the intersection of all the convex sets containing X . The convex hull of a finite set is called *polytope*.

The following result provides a characterization of the convex hull. By applying this characterization, it is easy to show that every polytope is compact.

Proposition 2.3 (Characterizing the convex hull) *Given a nonempty set $X \subset \mathbb{R}^n$, $\text{conv } X$ is the set of all the convex combinations of elements of X ; i.e.,*

$$\text{conv } X = \left\{ \sum_{i=1}^k \lambda_i x_i : x_i \in X, \lambda_i \geq 0, \sum_{i=1}^k \lambda_i = 1 \right\}.$$

Proof Let $Y := \left\{ \sum_{i=1}^k \lambda_i x_i : x_i \in X, \lambda_i \geq 0, \sum_{i=1}^k \lambda_i = 1 \right\}$. It is easy to prove that Y is a convex set containing X . Hence, we have $\text{conv } X \subset Y$.

Conversely, if $y \in Y$, as y is a convex combination of elements of X , it must belong to each convex set containing X . Therefore, $\text{conv } X \supset Y$, which completes the proof. \square

For a general set X , its closure, $\text{cl } X$, can be much bigger than its interior, $\text{int } X$ (e.g., $\text{int } \mathbb{Q} = \emptyset$ while $\text{cl } \mathbb{Q} = \mathbb{R}$). This is not the case of the convex sets, as the next result shows.

Proposition 2.4 (Interior and closure of convex sets) *If $C \subset \mathbb{R}^n$ is convex, then $\text{cl } C$ and $\text{int } C$ are also convex. Moreover, if $\text{int } C \neq \emptyset$, then $\text{cl } C = \text{cl int } C$.*

Proof Let $x, y \in \text{cl } C$. Then, there exist two sequences $\{x_k\}$ and $\{y_k\}$ contained in C such that $x_k \rightarrow x$ and $y_k \rightarrow y$. Given $\mu \in [0, 1]$, we have $(1 - \mu)x_k + \mu y_k \in C$ for all k . As $(1 - \mu)x_k + \mu y_k \rightarrow (1 - \mu)x + \mu y$, we conclude that $(1 - \mu)x + \mu y \in \text{cl } C$. Hence, $\text{cl } C$ is convex.

Let $x, y \in \text{int } C$ and $\mu \in [0, 1]$. Let $\varepsilon > 0$ be such that $x + \varepsilon \mathbb{B} \subset C$ and $y + \varepsilon \mathbb{B} \subset C$. Multiplying these inclusions by $1 - \mu$ and by μ , respectively, and adding them, we get $(1 - \mu)x + \mu y + \varepsilon \mathbb{B} \subset C$. Therefore, $(1 - \mu)x + \mu y \in \text{int } C$ and we can assert that $\text{int } C$ is convex (although possibly empty).

Assume now that $\text{int } C \neq \emptyset$. As $\text{int } C \subset C$, the inclusion $\text{cl } C \supset \text{cl int } C$ is trivial. To prove the reverse inclusion, let $x \in \text{cl } C$. Then, there exists a sequence $\{x_k\} \subset C$ such that $x_k \rightarrow x$. Moreover, as $\text{int } C \neq \emptyset$, there exists $y \in \text{int } C$. We shall prove that x can be obtained as a limit of points of $\text{int } C$. For that, we define the sequence

$$z_k := \left(1 - \frac{1}{k}\right)x_k + \frac{1}{k}y, \quad k = 1, 2, \dots$$

We shall see that $z_k \in \text{int } C$. Indeed, as $y \in \text{int } C$, there exists $\varepsilon > 0$ such that $y + \varepsilon \mathbb{B} \subset C$, so we have that

$$z_k + \frac{\varepsilon}{k} \mathbb{B} = \left(1 - \frac{1}{k}\right)x_k + \frac{1}{k}(y + \varepsilon \mathbb{B}) \subset \left(1 - \frac{1}{k}\right)x_k + \frac{1}{k}C \subset C,$$

where we have used the convexity of C in the last inclusion. Hence, $z_k \in \text{int } C$. As $z_k \rightarrow x$, we have proved that $x \in \text{cl int } C$. \square

Definition 2.5 A hyperplane *strictly separates* two sets when each of the sets is contained in a different open halfspace of the two determined by such a hyperplane.

Theorem 2.6 (Strict separation theorem) *Every closed convex set can be strictly separated from any external point.*

Proof Let C be a nonempty closed convex set and $y \notin C$. As the function $x \mapsto \|y - x\|^2$ is continuous on \mathbb{R}^n , it attains its minimum on C at a point c (applying Weierstrass theorem if C is bounded, and otherwise, the coercivity of such a function on C). Let $a := y - c$ and $\beta := a^T c$. Since

$$a^T y = a^T(y - c + c) = \|a\|^2 + a^T c,$$

we have $a^T y > \beta$. We shall begin by showing that $a^T x \leq \beta$, for all $x \in C$.

Let $x \in C$ and $0 < \lambda < 1$. As $(1 - \lambda)c + \lambda x \in C$ by the convexity of C , we have

$$\begin{aligned} \|c - y\|^2 &\leq \|(1 - \lambda)c + \lambda x - y\|^2 = \|(c - y) + \lambda(x - c)\|^2 \\ &= \|y - c\|^2 + \lambda^2\|x - c\|^2 + 2\lambda(c - y)^T(x - c). \end{aligned} \quad (2.1)$$

From (2.1) and $\lambda > 0$, it follows

$$\lambda\|x - c\|^2 - 2a^T(x - c) \geq 0. \quad (2.2)$$

Taking $\lim_{\lambda \searrow 0}$ in (2.2), we obtain $-2a^T(x - c) \geq 0$; that is, we deduce that $\beta = a^T c \geq a^T x$.

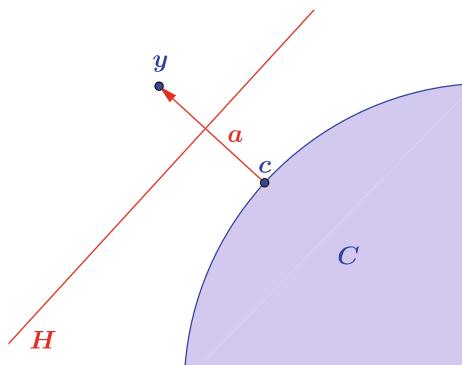
Finally, since

$$a^T y > b := \frac{a^T y + \beta}{2} > \beta \geq a^T x, \quad \forall x \in C,$$

we conclude that the hyperplane $H := \{x \in \mathbb{R}^n \mid a^T x = b\}$ strictly separates C and $\{y\}$ (see Fig. 2.1). \square

Definition 2.7 A hyperplane H supports the set X at a point $\bar{x} \in X$ if $\bar{x} \in H$ and X is contained in one of the two closed halfspaces determined by H .

Fig. 2.1 H strictly separates y from C



Remark 2.8 Observe that if $H = \{x \in \mathbb{R}^n \mid a^T x = b\}$ supports X at \bar{x} , one has (after multiplying both members of $a^T x = b$ by -1 if necessary), $a^T \bar{x} = b$ and $a^T x \geq b$, for all $x \in X$. Taking $y_k := \bar{x} - \frac{a}{k}$, we obtain $a^T y_k < b$, so we have that $\{y_k\} \subset \mathbb{R}^n \setminus X$, $k = 1, 2, \dots$, with $\lim_k y_k = \bar{x}$. As $\bar{x} \in X$, we must have $\bar{x} \in \text{bd } X$.

Theorem 2.9 (Supporting hyperplane theorem) *If C is a closed convex set and $\bar{x} \in \text{bd } C$, then there exists a hyperplane supporting C at \bar{x} .*

Proof As $\bar{x} \in \text{bd } X$, there exists a sequence $\{y_k\} \subset \mathbb{R}^n \setminus C$ such that $\lim_k y_k = \bar{x}$. By the separation theorem (Theorem 2.6), for each k there exists $a_k \in \mathbb{R}^n$, with $\|a_k\| = 1$, such that

$$a_k^T y_k > a_k^T x, \quad \forall x \in C. \quad (2.3)$$

As $\{a_k\}$ is contained in the unit sphere, which is compact, $\{a_k\}$ contains a convergent subsequence that we can suppose without loss of generality that it is $\{a_k\}$. Let $\lim_k a_k = a$, with $\|a\| = 1$. Taking limits in (2.3) when $k \rightarrow \infty$, one obtains

$$a^T \bar{x} \geq a^T x, \quad \forall x \in C.$$

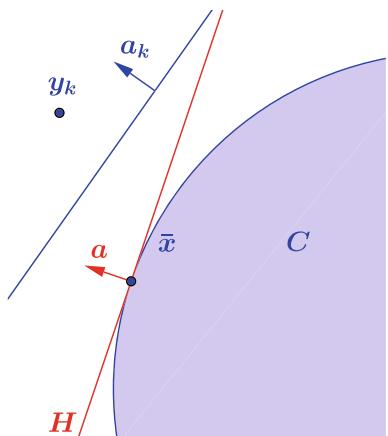
We conclude that $H = \{x \in \mathbb{R}^n \mid a^T x = a^T \bar{x}\}$ supports C at \bar{x} (see Fig. 2.2). \square

A particularly important class of convex sets is formed by the closed convex cones.

Definition 2.10 A set K such that $0_n \in K \subset \mathbb{R}^n$ is a *cone* if $\lambda x \in K$ for all $x \in K$ and all $\lambda \geq 0$.

Definition 2.11 The *conical (convex) hull* of a set $X \subset \mathbb{R}^n$ is the smallest convex cone containing X . It is denoted by $\text{cone } X$, and it is the intersection of all the convex cones containing X . Obviously, $\text{cone } \emptyset = \{0_n\}$.

Fig. 2.2 H supports C at \bar{x}



Proposition 2.12 (Characterizing the conical hull) *If $X \neq \emptyset$, then $\text{cone } X$ is the set of all the nonnegative linear combinations of points of X ; i.e.,*

$$\text{cone } X = \left\{ \sum_{i=1}^k \lambda_i x_i : x_i \in X, \lambda_i \geq 0 \right\}.$$

Proof The proof is similar to the one of Proposition 2.3, taking into account that if $y = \sum_{i=1}^k \lambda_i x_i$ with $x_i \in X$ and $\lambda_i \geq 0$ such that $L := \sum_{i=1}^k \lambda_i > 0$, then one has that

$$y = L \left(\sum_{i=1}^k \frac{\lambda_i}{L} x_i \right), \quad \text{where } \sum_{i=1}^k \frac{\lambda_i}{L} = 1.$$

Therefore, we have proved that y belongs to any convex cone containing X . \square

Next, we prove that any finitely generated convex cone is closed.

Proposition 2.13 (Closedness of the conical hull) *Let $X := \{x_1, \dots, x_k\} \subset \mathbb{R}^n$. Then, $\text{cone } X$ is closed.*

Proof By Proposition 2.12, we have that

$$\text{cone } X = \left\{ y \in \mathbb{R}^n : \sum_{i=1}^k \lambda_i x_i - y = 0, \lambda_i \geq 0 \forall i = 1, \dots, k \right\}.$$

Applying the Fourier elimination method k times to the previous system for eliminating the variables $\lambda_1, \dots, \lambda_k$, we obtain an homogeneous system involving only the variables y_1, \dots, y_k ; i.e., there exists a matrix A such that $\text{cone } X = \{y \in \mathbb{R}^n : Ay \leq 0\}$. In particular, from here we deduce that $\text{cone } X$ is closed. \square

We now illustrate the use of the *Fourier elimination method* for linear inequalities, which extends the well-known Gauss elimination method for linear equations, to get polyhedral representations of finitely generated convex cones and polytopes by means of a simple example. Hence, the closedness of finitely generated convex cones and polytopes can be shown by appealing to pure algebraic arguments.

Example 2.14 (a) We would like to express the convex cone generated by the vectors $(1, 0, 1)^T$, $(0, 1, 1)^T$, $(-1, 0, 0)^T$, $(1, -1, 0)^T$, and $(1, 1, 1)^T$ as the solution set of an homogeneous linear system in \mathbb{R}^3 . To this end, we must eliminate the parameters λ_i , $i = 1, \dots, 5$, from the linear system

$$\left\{ \begin{array}{l} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \lambda_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + \lambda_3 \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} + \lambda_4 \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} + \lambda_5 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \\ \lambda_i \geq 0, \quad i = 1, \dots, 5 \end{array} \right\}. \quad (2.4)$$

The first step consists on the Gauss elimination of three parameters, e.g., λ_1 , λ_2 , and λ_3 , from the subsystem formed by the three equations in (2.4), getting the reduced system

$$\left\{ \begin{array}{l} -x_2 + x_3 - \lambda_4 \geq 0 \\ x_2 + \lambda_4 - \lambda_5 \geq 0 \\ -x_1 - x_2 + x_3 + \lambda_5 \geq 0 \\ \lambda_4 \geq 0 \\ \lambda_5 \geq 0 \end{array} \right\}$$

or, equivalently,

$$\left\{ \begin{array}{l} \max\{-x_2 + \lambda_5, 0\} \leq \lambda_4 \leq -x_2 + x_3 \\ -x_1 - x_2 + x_3 + \lambda_5 \geq 0 \\ \lambda_5 \geq 0 \end{array} \right\}. \quad (2.5)$$

The second step, the Fourier elimination of λ_4 in (2.5), provides the system

$$\left\{ \begin{array}{l} -x_1 - x_2 + x_3 + \lambda_5 \geq 0 \\ \lambda_5 \leq x_3 \\ -x_2 + x_3 \geq 0 \\ \lambda_5 \geq 0 \end{array} \right\},$$

which is equivalent to

$$\left\{ \begin{array}{l} \max\{x_1 + x_2 - x_3, 0\} \leq \lambda_5 \leq x_3 \\ x_2 - x_3 \leq 0 \end{array} \right\}. \quad (2.6)$$

Finally, the Fourier elimination of λ_5 in (2.6) yields the aimed representation of the polyhedral cone, shown in Fig. 2.3:

$$\left\{ \begin{array}{l} x_1 + x_2 - 2x_3 \leq 0 \\ x_2 - x_3 \leq 0 \\ -x_3 \leq 0 \end{array} \right\}.$$

(b) To express the convex hull of the same vectors, $(1, 0, 1)^T$, $(0, 1, 1)^T$, $(-1, 0, 0)^T$, $(1, -1, 0)^T$, and $(1, 1, 1)^T$, as the solution set of a linear system in \mathbb{R}^3 we must eliminate λ_i , $i = 1, \dots, 5$, from

$$\left\{ \begin{array}{l} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \lambda_1 \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix} + \lambda_2 \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix} + \lambda_3 \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} + \lambda_4 \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} + \lambda_5 \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \\ \lambda_1 + \dots + \lambda_5 = 1 \\ \lambda_i \geq 0, i = 1, \dots, 5 \end{array} \right\}. \quad (2.7)$$

Eliminating via Gauss λ_1 , λ_2 , λ_3 , and λ_4 from the subsystem of (2.7) formed by the four equations, we get

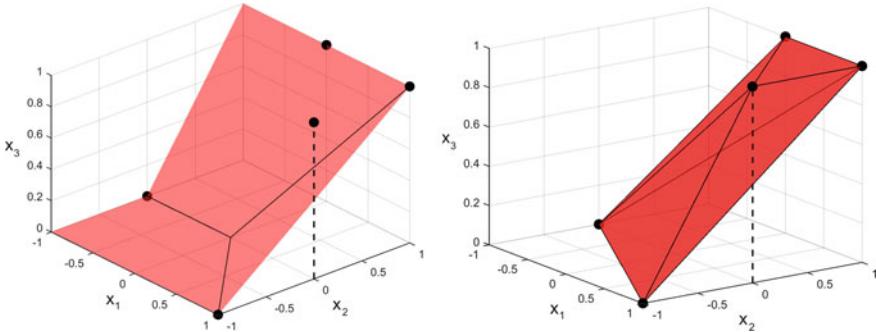


Fig. 2.3 Convex cone (left) and convex hull (right) generated by $(1, 0, 1)^T, (0, 1, 1)^T, (-1, 0, 0)^T, (1, -1, 0)^T$, and $(1, 1, 1)^T$

$$\begin{cases} -x_1 - 2x_2 + 3x_3 + \lambda_5 \geq 1 \\ x_1 + 2x_2 - 2x_3 - 2\lambda_5 \geq -1 \\ -x_1 - x_2 + x_3 + \lambda_5 \geq 0 \\ x_1 + x_2 - 2x_3 - \lambda_5 \geq -1 \\ \lambda_5 \geq 0 \end{cases}.$$

Eliminating now the remaining parameter λ_5 , one gets the aimed linear representation of the polytope, shown in Fig. 2.3:

$$\begin{cases} -x_1 \geq -1 \\ -x_3 \geq -1 \\ -x_1 - 2x_2 + 4x_3 \geq 1 \\ -x_2 + x_3 \geq 0 \\ x_1 + 2x_2 - 2x_3 \geq -1 \\ x_1 + x_2 - 2x_3 \geq -1 \end{cases}.$$

Now we shall obtain two useful consequences of the separation and the supporting hyperplane theorems.

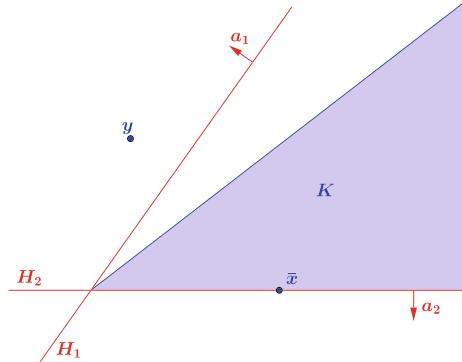
Corollary 2.15 (Separation theorem for convex cones) *If K is a closed convex cone and $y \notin K$, then there exists a vector $a \neq 0_n$ such that $a^T y > 0$ and $a^T x \leq 0$, for all $x \in K$.*

Proof By the separation theorem (Theorem 2.6), there exist $a \neq 0_n$ and $b \in \mathbb{R}$ such that $a^T y > b$ and $a^T x \leq b$, for all $x \in K$. We shall prove that we can take $b = 0$ (see Fig. 2.4).

On the one hand, as $0_n \in K$, it must hold $a^T 0_n = 0 \leq b < a^T y$. On the other hand, given $x \in K$, $\lambda x \in K$ for all $\lambda > 0$, so we have that $a^T (\lambda x) \leq b$ or, equivalently, $a^T x \leq b/\lambda$. Taking $\lambda \rightarrow +\infty$, we obtain $a^T x \leq 0$. \square

Corollary 2.16 *If K is a closed convex cone and $\bar{x} \in \text{bd } K$, there exists a hyperplane containing the origin and supporting K at \bar{x} .*

Fig. 2.4 H_1 strictly separates y from K and H_2 supports K at \bar{x}



Proof As $\bar{x} \in \text{bd } K$, by the supporting hyperplane theorem (Theorem 2.9), there exist $a \neq 0_n$ and $b \in \mathbb{R}$ such that $a^T x \geq b$ for all $x \in K$ and $a^T \bar{x} = b$. It will be sufficient to prove that $b = a^T 0_n = 0$ (see Fig. 2.4).

On the one hand, as $0_n \in K$, one has $0 = a^T 0_n \geq b$. On the other hand, as $a^T(\lambda \bar{x}) \geq b$ for all $\lambda > 0$, $b = a^T \bar{x} \geq b/\lambda$. Taking $\lambda \rightarrow +\infty$, we obtain $b \geq 0$. We conclude that $b = 0$. \square

2.2 Convex Functions of One Variable

In this section, unless otherwise indicated, I represents a proper interval in \mathbb{R} (i.e., an interval containing more than one point). Convex functions are those whose graph arcs are below the corresponding chords. The precise definition is as follows.

Definition 2.17 The function $f : I \rightarrow \mathbb{R}$ is *convex* (on I) if

$$f((1 - \mu)x + \mu y) \leq (1 - \mu)f(x) + \mu f(y), \quad \forall x, y \in I, \forall \mu \in]0, 1[.$$

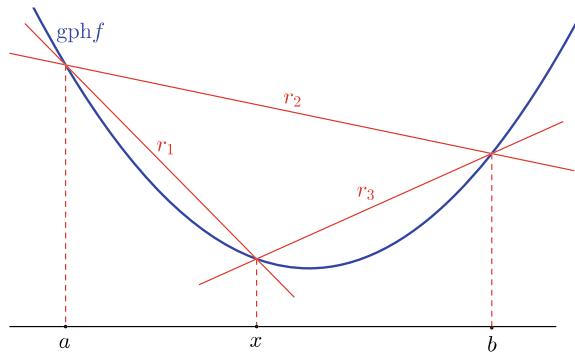
The function f is *concave* when $-f$ is convex.

Recall that if $a \in I$ and $a \neq \max I$, the *right derivative* of f at a is $\lim_{x \searrow a} \frac{f(x) - f(a)}{x - a}$, whenever this limit exists (it can be infinite). It is denoted by $f'_+(a)$. If f is right derivable at a and $f'_+(a) \in \mathbb{R}$, the *right-hand tangent* to $\text{gph } f$ at $(a, f(a))$ is $y = f(a) + f'_+(a)(x - a)$. In an analogous way, we define the *left derivative* of f at $a \neq \min I$, denoted by $f'_-(a)$, and the *left-hand tangent* to $\text{gph } f$ at $(a, f(a))$.

Obviously, $f : I \rightarrow \mathbb{R}$ is *differentiable* (or *derivable*) at $a \in \text{int } I$ if and only if both lateral derivatives exist at a and $f'_+(a) = f'_-(a) \in \mathbb{R}$. This real number is the *derivative* of f at a , $f'(a) = \lim_{x \rightarrow a} \frac{f(x) - f(a)}{x - a}$. In such a case, the *tangent* to $\text{gph } f$ at $(a, f(a))$ is $y = f(a) + f'(a)(x - a)$. Observe that the function $f(x) = \sqrt[3]{x}$ lacks of derivative at 0 although $f'_+(0) = f'_-(0)$, since that common value is $+\infty$.

Fig. 2.5 Illustration of (i):

$m_1 \leq m_2 \leq m_3$,
where m_i is the
slope of the line
 $r_i, i = 1, 2, 3$



The following result, proved in 1983, collects different properties of convex functions of one variable regarding the slopes of chords and lateral tangents of their graphs.

Theorem 2.18 (Stolz Theorem) *Let $f : I \rightarrow \mathbb{R}$ be a convex function. Then, the following statements hold:*

(i) If $a, b \in I$ and $a < x < b$, then

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

(ii) There exist f'_+ and f'_- at every point of I in which they can exist. Moreover, f'_+ and f'_- are finite on $\text{int } I$, so f is continuous on $\text{int } I$.

(iii) $f'_- \leq f'_+$ on $\text{int } I$.

(iv) If $x, y \in I$ and $x < y$, then

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

(v) f'_- and f'_+ are nondecreasing on $\text{int } I$.

(vi) $\text{gph } f$ is above the lateral tangents at $(x, f(x))^T$ for all $x \in \text{int } I$.

Proof The convexity of f will be only used to prove (i), which is the property that allows proving the other five statements.

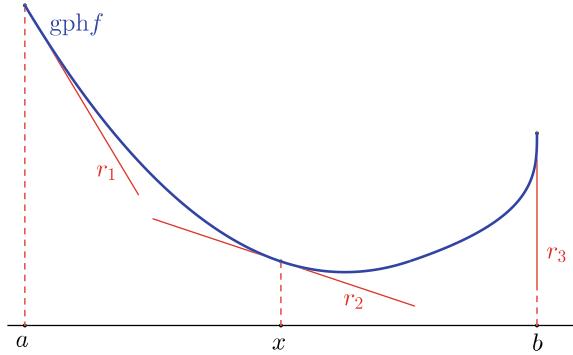
(i) Figure 2.5, where m_i represents the slope of $r_i, i = 1, 2, 3$, illustrates statement (i). We shall prove the first inequality.

Any point $x \in]a, b[$ can be expressed as a convex combination of the extreme points of the segment, and finding the value of μ in $x = (1 - \mu)a + \mu b$, we have $\mu = \frac{x-a}{b-a}$. Then, by the convexity of f ,

$$f(x) \leq (1 - \mu)f(a) + \mu f(b) = f(a) + \left(\frac{x-a}{b-a}\right)(f(b) - f(a)). \quad (2.8)$$

Fig. 2.6 Illustration of (ii):

$f'_+(a) = m_1$,
 $f'_-(x) = f'_+(x) = m_2$,
and $f'_-(b) = +\infty$



Subtracting $f(a)$ and dividing by $x - a > 0$ both members of (2.8), we obtain

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

The second inequality can be analogously proved, subtracting $f(b)$ and dividing by $x - b < 0$.

(ii) Figure 2.6 illustrates statement (ii). In this figure, the lateral derivatives coincide on $\text{int } I$, but it is easy to see that such a thing does not always occur (see Fig. 2.9).

We associate to $c \in I$ the incremental quotient $g_c : I \setminus \{c\} \rightarrow \mathbb{R}$ such that $g_c(x) = \frac{f(x) - f(c)}{x - c}$.

Let $c \in I$ be such that $c \neq \max I$. We shall prove that there exists $f'_+(c)$, i.e., $\lim_{x \searrow c} g_c(x)$. For that, it is sufficient to show that g_c is nondecreasing on $\{x \in I : x > c\}$, since in such a case it holds

$$\lim_{x \searrow c} g_c(x) = \inf\{g_c(x) : x \in I, x > c\} \in \mathbb{R} \cup \{-\infty\}.$$

Let $x_1, x_2 \in I$ be such that $c < x_1 < x_2$. As we can see in Fig. 2.7, by applying (i) to f one obtains

$$g_c(x_1) = \frac{f(x_1) - f(c)}{x_1 - c} \leq \frac{f(x_2) - f(c)}{x_2 - c} = g_c(x_2).$$

We have proved that

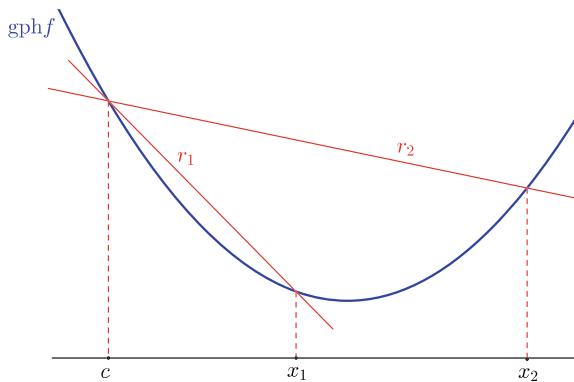
$$f'_+(c) = \inf \left\{ \frac{f(x) - f(c)}{x - c} : x \in I, x > c \right\} \in \mathbb{R} \cup \{-\infty\}. \quad (2.9)$$

Analogously, one shows that

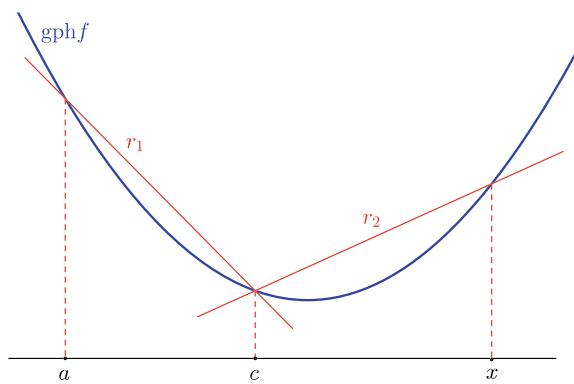
$$f'_-(c) = \sup \left\{ \frac{f(x) - f(c)}{x - c} : x \in I, x < c \right\} \in \mathbb{R} \cup \{+\infty\}. \quad (2.10)$$

Fig. 2.7 Meaning of g_c :

$g_c(x_1) = m_1$,
 $g_c(x_2) = m_2$,
with $m_1 \leq m_2$

**Fig. 2.8** $\frac{f(c) - f(a)}{c - a} = m_1$

and $g_c(x) = m_2$,
with $m_1 \leq m_2$



Now, suppose that $c \in \text{int } I$. We shall prove that $f'_+(c) \in \mathbb{R}$, i.e., $\lim_{x \searrow c} g_c(x)$ is finite. For that, it is sufficient to show that g_c is bounded from below on $\{x \in I : x > c\}$. Let $x \in I$ be such that $x > c$. We have $a < c < x$. Applying (i) again (see Fig. 2.8), we have:

$$\frac{f(c) - f(a)}{c - a} \leq \frac{f(x) - f(c)}{x - c} = g_c(x).$$

Then, $f'_+(c) = \lim_{x \searrow c} g_c(x)$ exists and is finite. In the same way, one shows that there exists $f'_-(c)$ when $c \in I \setminus \{\min I\}$ and that it is finite when $c \in \text{int } I$. Let $c \in \text{int } I$. As $\lim_{x \searrow c} \frac{f(x) - f(c)}{x - c} \in \mathbb{R}$ and $\lim_{x \searrow c} (x - c) = 0$, we have that $\lim_{x \searrow c} [f(x) - f(c)] = 0$, i.e., $\lim_{x \searrow c} f(x) = f(c)$. The same thing happens on the left-hand side of c , so f is continuous at c .

(iii) Figure 2.9 illustrates statement (iii).

Let $c \in \text{int } I$. Let $\delta > 0$ be such that $c - \delta, c + \delta \in I$. Let $0 < h < \delta$. Applying (i) to $c - h < c < c + h$ (see Fig. 2.10), we have

Fig. 2.9 Illustration of (iii):
 $f'_-(c) = m_1$ and
 $f'_+(c) = m_2$, with
 $m_1 \leq m_2$

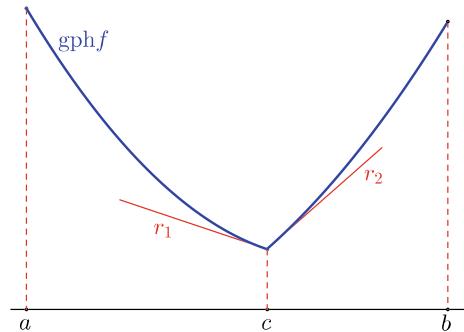


Fig. 2.10 $\frac{f(c) - f(c-h)}{h} = m_1$,
 $\frac{f(c+h) - f(c)}{h} = m_2$,
and $m_1 \leq m_2$

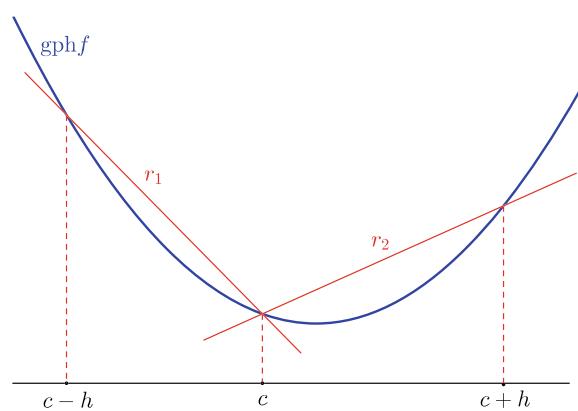
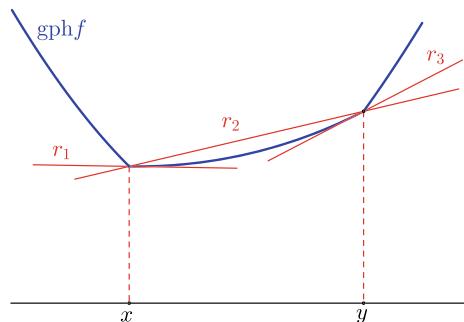


Fig. 2.11 Illustration of (iv):
 $m_1 \leq m_2 \leq m_3$,
where $f'_+(x) = m_1$,
 $\frac{f(y) - f(x)}{y - x} = m_2$
and $f'_-(y) = m_3$



$$\frac{f(c) - f(c-h)}{h} \leq \frac{f(c+h) - f(c)}{h}. \quad (2.11)$$

Taking $h \searrow 0$ in both members of (2.11), one obtains $f'_-(c) \leq f'_+(c)$.

(iv) Figure 2.11 illustrates this proposition.

Let z be such that $x < z < y$. Applying the double inequality of (i) to $x < z < y$ (see Fig. 2.12), we have

Fig. 2.12 $m_1 \leq m_2 \leq m_3$,
 $\frac{f(z) - f(x)}{z - x} = m_1$,
 $\frac{f(y) - f(x)}{y - x} = m_2$,
 $\frac{f(y) - f(z)}{y - z} = m_3$

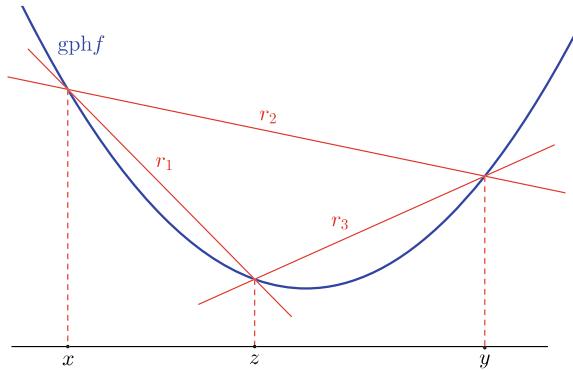
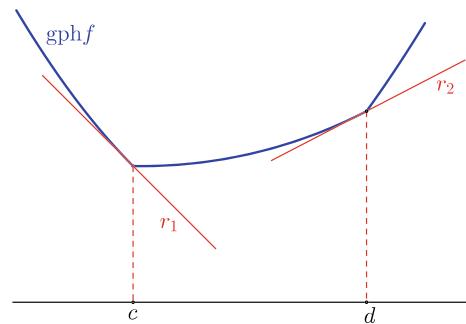


Fig. 2.13 Illustration of (v):

$m_1 \leq m_2$, with
 $f'_-(c) = m_1$, and
 $f'_-(d) = m_2$



$$\frac{f(z) - f(x)}{z - x} \leq \frac{f(y) - f(x)}{y - x} \leq \frac{f(y) - f(z)}{y - z}.$$

By taking $z \searrow x$ and $z \nearrow y$ in the first and in the second inequality, we obtain

$$f'_+(x) \leq \frac{f(y) - f(x)}{y - x} \leq f'_-(y),$$

completing thereby the proof of statement (iv).

(v) This assertion is proved combining (iii) and (iv). Figures 2.13 and 2.14 illustrate statement (v).

(vi) Figure 2.15 illustrates statement (vi).

Let $c \in \text{int } I$. By (2.9), we have $f'_+(c) \leq \frac{f(x) - f(c)}{x - c}$ for all $x \in I$ such that $x > c$, whereas, by (2.10) and (iii), $\frac{f(x) - f(c)}{x - c} \leq f'_-(c) \leq f'_+(c)$ for all $x \in I$ such that $x < c$. Thus, we obtain

$$f(x) \geq f(c) + f'_+(c)(x - c), \quad \forall x \in I. \quad (2.12)$$

From (2.12), we deduce that $\text{gph } f$ is above the right-hand tangent at $(c, f(c))^T$. The similar result for the left-hand tangent can be analogously proved. \square

Next, we recall the definition of Lipschitz continuity.

Fig. 2.14 $m_1 \leq m_2$, with
 $f'_+(c) = m_1$, and
 $f'_+(d) = m_2$

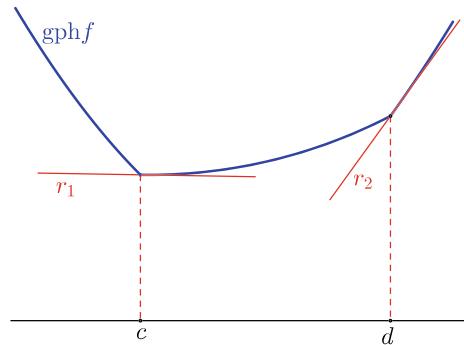
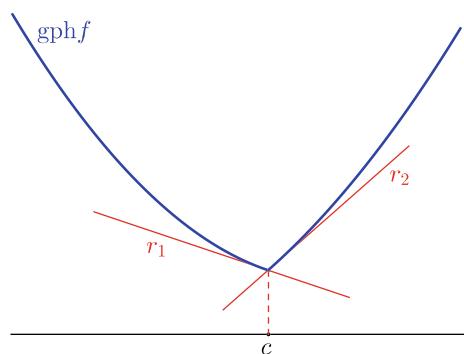


Fig. 2.15 Illustration of (vi):
 r_1 and r_2 are
the lines
 $y = f(c) + f'_-(c)(x - c)$
and
 $y = f(c) + f'_+(c)(x - c)$,
respectively



Definition 2.19 A function $f : I \rightarrow \mathbb{R}$ is *Lipschitz continuous* with constant $L \geq 0$ at the point $a \in I$ if

$$|f(x) - f(a)| \leq L|x - a|, \quad \forall x \in I, \quad (2.13)$$

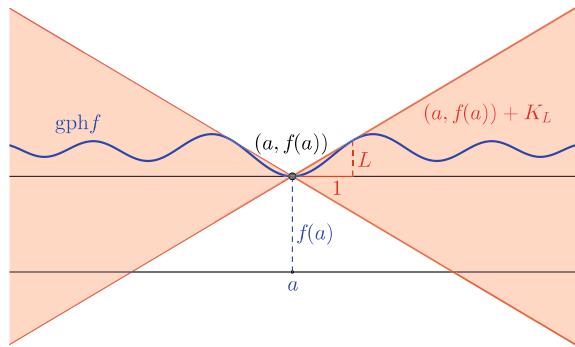
(with $L = 0$ if and only if f is constant on I). The function f is said to be *Lipschitz continuous* with constant L on I if (2.13) holds at any point of I .

This condition, which obviously implies the continuity of f at a , is equivalent to

$$\left| \frac{f(x) - f(a)}{x - a} \right| \leq L, \quad \forall x \in I \setminus \{a\},$$

which admits the following geometrical interpretation: L is an upper bound for the absolute value of the chord slopes of $\text{gph } f$ having $(a, f(a))^T$ as one of their extreme points. In other words, $\text{gph } f \subset (a, f(a))^T + K_L$, where $K_L := \{(x, y)^T \in \mathbb{R}^2 : |y| \leq L|x|\}$ (a double cone containing the origin); see Fig. 2.16. Finally, Lipschitz continuity with constant L on I is equivalent to the inclusion $\text{gph } f \subset (x, f(x))^T + K_L$ for all $x \in I$. Observe that $f(x) = \sqrt[3]{x}$ is not Lipschitz continuous at 0.

Fig. 2.16 Interpretation of Lipschitz continuity of f at a



Example 2.20 The function $f(x) = x^2$ is convex and Lipschitz continuous on the intervals of the form $[a, b]$, with $a < b$, but not on its domain \mathbb{R} . The details are left to the reader.

Proposition 2.21 (Lipschitz continuity) *Let I be an open interval, and let $f : I \rightarrow \mathbb{R}$ be a convex function. Then, f is Lipschitz continuous on any interval $[a, b] \subset I$, $a < b$, with constant*

$$L := \max\{|f'_+(a)|, |f'_-(b)|\}. \quad (2.14)$$

Proof Let $a, b \in I$ and $x, y \in [a, b]$, with $x < y$. Taking L as in (2.14), by Stolz Theorem 2.18(ii), one has that $L \in \mathbb{R}_+$. Combining this together with statements (iv) and (v) of Stolz Theorem 2.18, we obtain

$$-L \leq f'_+(a) \leq f'_+(x) \leq \frac{f(y) - f(x)}{y - x} \leq f'_-(y) \leq f'_-(b) \leq L,$$

so we have $\left| \frac{f(y) - f(x)}{y - x} \right| \leq L$, i.e., $|f(y) - f(x)| \leq L|y - x|$. As x and y are interchangeable, we conclude that, if $x, y \in [a, b]$, it holds

$$|f(y) - f(x)| \leq L|y - x|.$$

This implies Lipschitz continuity of f on $[a, b]$ with constant L , as claimed. \square

From Proposition 2.21, we could deduce that f is continuous on I , as we already knew by Theorem 2.18(ii), since I can be expressed as a countable and expansive union of intervals of the form $[a, b] \subset I$.

Example 2.22 The function $f(x) = -\sqrt{1 - x^2}$ for all $x \in]-1, 1[$ and $f(\pm 1) = 1$ is convex, but not Lipschitz continuous on its domain $[-1, 1]$, nor on its interior $] -1, 1[$, although it is on any interval $[a, b]$ such that $-1 < a < b < 1$, being the smallest Lipschitz constant $L := \frac{z}{\sqrt{1-z^2}}$, where $z = \max\{|a|, |b|\}$, as we can see in Fig. 2.17.

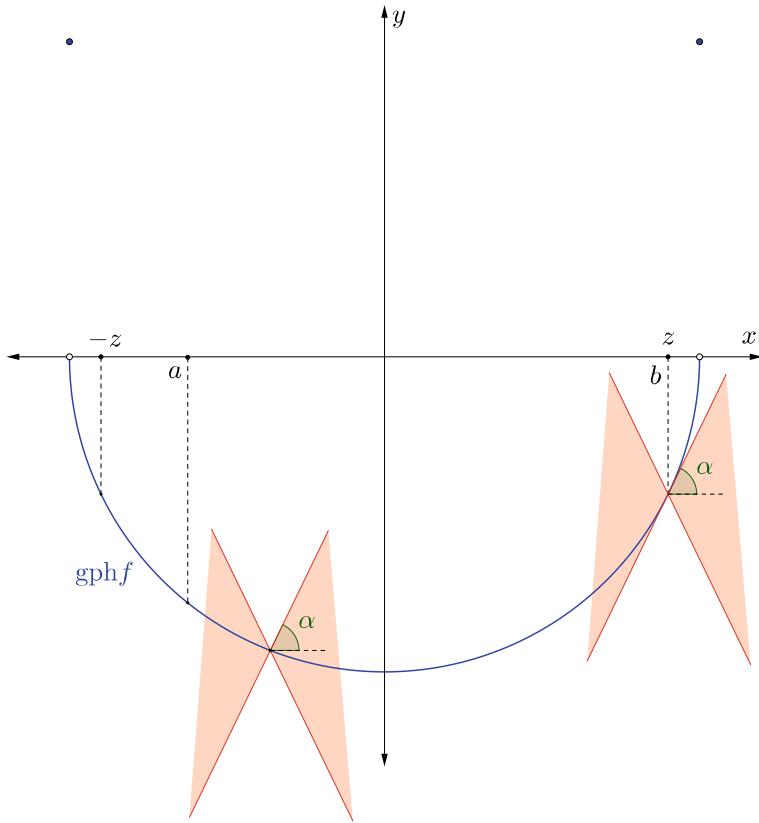


Fig. 2.17 The function f of Example 2.22 is Lipschitz continuous on $[a, b]$ with constant $L = \tan \alpha$

We shall finish this section by characterizing the convexity of smooth functions of one variable.

Theorem 2.23 (Characterization of differentiable convex functions) *Let $f : I \rightarrow \mathbb{R}$ be differentiable on I , where I is an open interval in \mathbb{R} . Then, f is convex if and only if f' is nondecreasing on I .*

Proof The direct implication is a straightforward consequence of Stolz Theorem 2.18, part (v).

Let us suppose that f' is nondecreasing on I . Let $a, b \in I$, $a < b$, and let $\mu \in]0, 1[$. Let $c := (1 - \mu)a + \mu b$. By applying the mean value theorem to f on the intervals $[a, c]$ and $[c, b]$, we can assert the existence of $x_1 \in]a, c[$ and $x_2 \in]c, b[$ such that

$$f(c) - f(a) = f'(x_1)(c - a); \quad (2.15)$$

$$f(b) - f(c) = f'(x_2)(b - c). \quad (2.16)$$

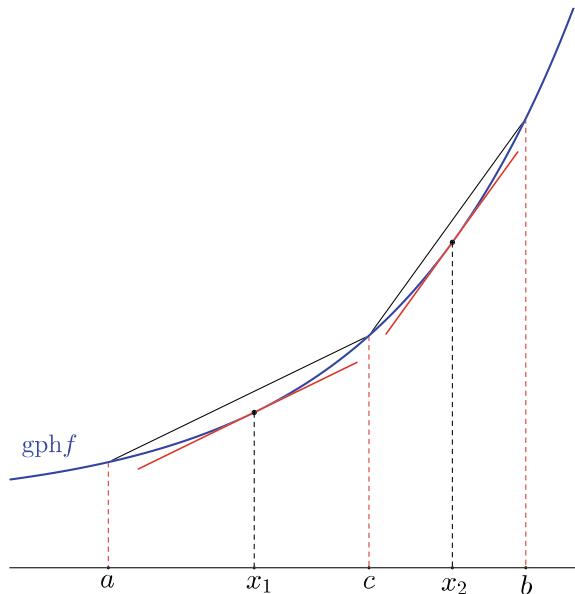
Fig. 2.18 $f'(x_1) \leq f'(x_2)$,

with

$$f'(x_1) = \frac{f(c)-f(a)}{c-a}$$

and

$$f'(x_2) = \frac{f(b)-f(c)}{b-c}$$



From (2.15) and (2.16), observing that $x_1 < x_2$ and that f' is nondecreasing on I , we deduce that

$$\frac{f(c)-f(a)}{c-a} = f'(x_1) \leq f'(x_2) = \frac{f(b)-f(c)}{b-c};$$

see Fig. 2.18.

Therefore,

$$\left(\frac{1}{c-a} + \frac{1}{b-c} \right) f(c) \leq \frac{f(a)}{c-a} + \frac{f(b)}{b-c},$$

i.e.,

$$f((1-\mu)a + \mu b) \leq (1-\mu)f(a) + \mu f(b),$$

so we have that f is convex. \square

Theorem 2.24 (Characterization of twice differentiable convex functions) *Let $f : I \rightarrow \mathbb{R}$ be twice differentiable on I , where I is an open interval. Then, f is convex if and only if f'' is nonnegative on I .*

Proof This is a consequence of Theorem 2.23 and the characterization of nondecreasing differentiable functions by means of the sign of the derivative. \square

Let $\emptyset \neq I \subset \mathbb{R}$ be an interval. If $f : \text{cl } I \rightarrow \mathbb{R}$ is continuous on $\text{cl } I$ and is convex on I , then f is convex on $\text{cl } I$, since the points of $\text{cl } I$ are the limit of sequences contained in I . If $f, g : \text{cl } I \rightarrow \mathbb{R}$ are such that f is convex, $g(x) = f(x)$ for all

$x \in I$ and $g(x) \geq f(x)$ for all $x \in \text{bd } I$, then g is convex on $\text{cl } I$. Indeed, given $\mu \in]0, 1[$ and $x, y \in \text{cl } I$, $x \neq y$, one has $(1 - \mu)x + \mu y \in \text{int } I$. Hence,

$$\begin{aligned} g((1 - \mu)x + \mu y) &= f((1 - \mu)x + \mu y) \leq (1 - \mu)f(x) + \mu f(y) \\ &\leq (1 - \mu)g(x) + \mu g(y). \end{aligned}$$

Example 2.25 By applying Theorem 2.24, we can prove analytically that the function in Example 2.22 is convex. Indeed, as $f''(x) > 0$ for all $x \in]-1, 1[$, the restriction of f to $] -1, 1[$, $f|_{]-1,1[}$, is convex, as well as its continuous extension to $[-1, 1]$ (the function $x \mapsto -\sqrt{1 - x^2}$) and, finally, the function f that results of adding one unit to its value at ± 1 .

2.3 Convex Functions of Several Variables

Definition 2.26 A function $f : C \rightarrow \mathbb{R}$, where $C \subset \mathbb{R}^n$ is a nonempty convex set, is *convex* (on C) if

$$f((1 - \mu)x + \mu y) \leq (1 - \mu)f(x) + \mu f(y), \quad \forall x, y \in C, \forall \mu \in]0, 1[.$$

Example 2.27 It is easy to show that the sublevel sets of convex functions are convex. The function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $f(x) = -\exp(-\|x\|^2)$ proves that the converse statement is not true.

Convex sets and convex functions are linked via the notion of epigraph (the prefix “epi” comes from the Greek and means “above”).

Definition 2.28 Let $f : C \rightarrow \mathbb{R}$, where $C \subset \mathbb{R}^n$ is a nonempty convex set. The *epigraph* of the function f is the set in \mathbb{R}^{n+1} given by

$$\text{epi } f := \left\{ \begin{pmatrix} x \\ \alpha \end{pmatrix} \in \mathbb{R}^{n+1} : x \in C, f(x) \leq \alpha \right\}.$$

It is straightforward to prove that if C is closed and $f : C \rightarrow \mathbb{R}$ is continuous, then $\text{epi } f$ is closed.

Proposition 2.29 (Epigraphical characterization of convexity) *Let $f : C \rightarrow \mathbb{R}$, where $C \subset \mathbb{R}^n$ is a nonempty convex set. The function f is convex if and only if the set $\text{epi } f$ is convex.*

Proof Suppose first that f is convex. Choose any $\begin{pmatrix} x \\ \alpha \end{pmatrix}, \begin{pmatrix} y \\ \beta \end{pmatrix} \in \text{epi } f$ and $\mu \in]0, 1[$. Since f is convex,

$$f((1 - \mu)x + \mu y) \leq (1 - \mu)f(x) + \mu f(y) \leq (1 - \mu)\alpha + \mu\beta,$$

which implies that $(1 - \mu)(\frac{x}{\alpha}) + \mu(\frac{y}{\beta}) \in \text{epi } f$.

Conversely, assume that $\text{epi } f$ is convex. Pick any $x, y \in C$ and $\mu \in]0, 1[$. Then, since $(\frac{x}{f(x)}, \frac{y}{f(y)}) \in \text{epi } f$, one has $(1 - \mu)(\frac{x}{f(x)}) + \mu(\frac{y}{f(y)}) \in \text{epi } f$, which implies convexity of the function f . \square

The importance of convex functions in optimization arises from the following elementary result.

Proposition 2.30 (Local minima are global) *Let $C \subset \mathbb{R}^n$ be a convex set and $f : C \rightarrow \mathbb{R}$ be a convex function. If \bar{x} is a local minimum of f on C , then it is a global minimum of f on C .*

Proof If \bar{x} is not a global minimum there exists another point $\tilde{x} \in C$ such that $f(\tilde{x}) < f(\bar{x})$. For all $\mu \in]0, 1[$, it holds

$$f((1 - \mu)\bar{x} + \mu\tilde{x}) \leq (1 - \mu)f(\bar{x}) + \mu f(\tilde{x}) < f(\bar{x}),$$

so any neighborhood of \bar{x} contains points where f has a lower value than $f(\bar{x})$. Therefore, \bar{x} is not a local minimum. \square

The next proposition shows how to generate convex functions from other convex functions. Subsequently, we will characterize smooth convex functions.

Proposition 2.31 (Generation of convex functions) *Let $C \subset \mathbb{R}^n$ be a nonempty convex set. The following statements hold:*

- (i) *If $f : C \rightarrow \mathbb{R}$ is convex and $\alpha \in \mathbb{R}_+$, then αf is convex.*
- (ii) *If $f, g : C \rightarrow \mathbb{R}$ are convex, then $f + g$ is convex.*
- (iii) *If $f_k : C \rightarrow \mathbb{R}$ is convex for all $k \in \mathbb{N}$ and $\lim_{k \rightarrow \infty} f_k = f$ pointwise on C , then f is convex.*
- (iv) *If $f_i : C \rightarrow \mathbb{R}$ is convex for all $i \in I$ and $\{f_i(x) : i \in I\}$ is upper bounded for all $x \in C$, then $\sup_{i \in I} f_i$ is convex.*
- (v) *If $f : C \rightarrow \mathbb{R}$ and $h : I \rightarrow \mathbb{R}$, with $f(C) \subset I$ (an interval on \mathbb{R}), are convex and h is nondecreasing, then $h \circ f : C \rightarrow \mathbb{R}$ is convex.*
- (vi) *If $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex and $g : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is an affine function, then $f \circ g : \mathbb{R}^m \rightarrow \mathbb{R}$ is convex.*

Proof (i), (ii), (iii), and (vi) are straightforward consequences of the definition.

(iv) Let us denote $h := \sup_{i \in I} f_i$. Under the hypothesis, the function h is defined on C , i.e., $h : C \rightarrow \mathbb{R}$. Let $i \in I$. Given $x, y \in C$ and $\mu \in]0, 1[$, one has

$$f_i((1 - \mu)x + \mu y) \leq (1 - \mu)f_i(x) + \mu f_i(y) \leq (1 - \mu)h(x) + \mu h(y).$$

Then, $h((1 - \mu)x + \mu y) \leq (1 - \mu)h(x) + \mu h(y)$.

(v) Let $x, y \in \mathbb{R}^n$ and $\mu \in]0, 1[$. Then

$$\begin{aligned} (h \circ f)((1 - \mu)x + \mu y) &= h(f((1 - \mu)x + \mu y)) \leq h((1 - \mu)f(x) + \mu f(y)) \\ &\leq (1 - \mu)(h \circ f)(x) + \mu(h \circ f)(y), \end{aligned}$$

completing thereby the proof. \square

Example 2.32 Let $f : \mathbb{B} \rightarrow \mathbb{R}$ be such that $f(x) = -\sqrt{1 - \|x\|^2}$. By Proposition 2.31(v) and decomposing f as $h \circ g$, where $h(t) = -\sqrt{1 - t}$ and $g(x) = \|x\|^2$, one shows that f is convex.

As a consequence of the supporting hyperplane theorem, there exists a nonvertical hyperplane supporting $\text{cl epi } f$ at any point in the interior of the domain of a convex function.

Proposition 2.33 (Existence of nonvertical supporting hyperplanes) *Let C be a convex set in \mathbb{R}^n with nonempty interior, and let $f : C \rightarrow \mathbb{R}$ be a convex function. Then, for any $\bar{x} \in \text{int } C$, there exists a nonvertical hyperplane supporting $\text{cl epi } f$ at $(\frac{\bar{x}}{f(\bar{x})})$. Furthermore, if f is differentiable at \bar{x} , then one has*

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x}), \quad \forall x \in C, \quad (2.17)$$

and the nonvertical supporting hyperplane is unique and defined by the right-hand side of (2.17).

Proof Pick any $\bar{x} \in C$. Clearly, one has that $(\frac{\bar{x}}{f(\bar{x})}) \in \text{bd epi } f$. By Propositions 2.4 and 2.29, $\text{cl epi } f$ is a closed convex set in \mathbb{R}^{n+1} . Then, by Theorem 2.9, there exists a hyperplane supporting $\text{cl epi } f$ at $(\frac{\bar{x}}{f(\bar{x})})$, which entails the existence of $(\frac{a}{\gamma}) \in \mathbb{R}^{n+1} \setminus \{0_{n+1}\}$ and $b \in \mathbb{R}$ such that

$$a^T \bar{x} + \gamma f(\bar{x}) = b \quad \text{and} \quad a^T x + \gamma y \geq b, \quad \forall (\frac{x}{y}) \in \text{cl epi } f. \quad (2.18)$$

To conclude the proof, we need to show that $\gamma > 0$. Applying (2.18) to $(\frac{\bar{x}}{1+f(\bar{x})})$, we obtain $\gamma \geq 0$. Now, we shall prove that $\gamma \neq 0$. This is clear when $a = 0_n$, because $(\frac{a}{\gamma}) \neq 0_{n+1}$. When $a \neq 0_n$, if $\gamma = 0$, then (2.18) implies in particular that $a^T \bar{x} = b$ and $a^T x \geq b$ for all $x \in C$. Since $\bar{x} \in \text{int } C$, there is $\varepsilon > 0$ such that $\bar{x} + \varepsilon \mathbb{B} \subset C$. Then, $\bar{x} - \varepsilon \frac{a}{\|a\|} \in C$, which implies

$$b \leq a^T \left(\bar{x} - \varepsilon \frac{a}{\|a\|} \right) = a^T \bar{x} - \varepsilon \|a\| = b - \varepsilon \|a\|,$$

a contradiction, since $a \neq 0_n$. Therefore, one must have $\gamma \neq 0$, as claimed.

Now suppose that f is differentiable at \bar{x} , and suppose that (2.18) holds for some $(\frac{a}{\gamma})$ with $\gamma > 0$. Hence,

$$\frac{1}{\gamma} a^T (x - \bar{x}) + (f(x) - f(\bar{x})) \geq 0, \quad \forall x \in C. \quad (2.19)$$

Consider the sequence $x_k := \bar{x} - \frac{1}{k} \left(\frac{a}{\gamma} + \nabla f(\bar{x}) \right)$, $k = 1, 2, \dots$. Then, $x_k \in C$ for all k sufficiently large, and we have

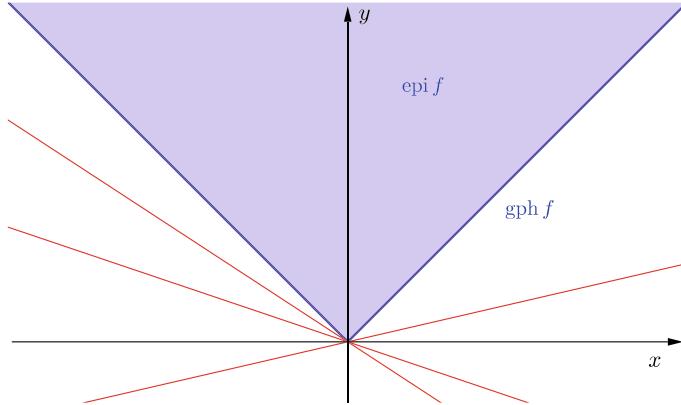


Fig. 2.19 The hyperplane supporting the epigraph at the origin is not unique

$$\begin{aligned}
 0 &\leq \frac{1}{\gamma} a^T (x_k - \bar{x}) + f(x_k) - f(\bar{x}) \\
 &= \left(\frac{a}{\gamma} + \nabla f(\bar{x}) \right)^T (x_k - \bar{x}) + f(x_k) - f(\bar{x}) - \nabla f(\bar{x})^T (x_k - \bar{x}) \\
 &= -\frac{1}{k} \left\| \frac{a}{\gamma} + \nabla f(\bar{x}) \right\|^2 + f(x_k) - f(\bar{x}) - \nabla f(\bar{x})^T (x_k - \bar{x}).
 \end{aligned}$$

Dividing the latter inequality by $\|x_k - \bar{x}\| = \frac{1}{k} \left\| \frac{a}{\gamma} + \nabla f(\bar{x}) \right\|$, we get

$$-\left\| \frac{a}{\gamma} + \nabla f(\bar{x}) \right\| + \frac{f(x_k) - f(\bar{x}) - \nabla f(\bar{x})^T (x_k - \bar{x})}{\|x_k - \bar{x}\|} \geq 0,$$

and letting $k \rightarrow \infty$, because of the definition of differentiability at \bar{x} (see (1.6)), we deduce that $\frac{a}{\gamma} = -\nabla f(\bar{x})$. Therefore, by (2.19), we have that (2.17) holds, which completes the proof. \square

Example 2.34 The continuous and convex function $f : \mathbb{R} \rightarrow \mathbb{R}$ given by $f(x) = |x|$ is differentiable at every point except $\bar{x} = 0$. By Proposition 2.33, there exists a nonvertical hyperplane supporting $\text{epi } f$ at every point in the graph. As shown in Fig. 2.19, the hyperplane is not unique at the origin. The same is true for $f : \mathbb{R}^n \rightarrow \mathbb{R}$ given by $f(x) = \|x\|$. The epigraph of the function $f(x) = -\sqrt{1 - x^2}$ in Example 2.22 has only vertical supporting hyperplanes at $(\pm 1, 0)^T$.

Let C be a nonempty convex set in \mathbb{R}^n . For each pair $x, d \in \mathbb{R}^n$ (interpreted as a point and a direction), we consider an affine function $g_{x,d} : \mathbb{R} \rightarrow \mathbb{R}^n$ such that $g_{x,d}(t) := x + td$. As C is convex,

$$I_{x,d} := \{t \in \mathbb{R} : x + td \in C\} = g_{x,d}^{-1}(C)$$

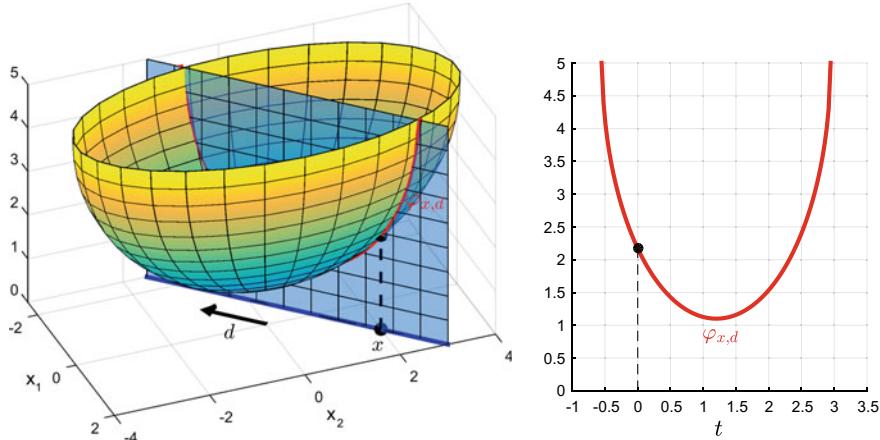


Fig. 2.20 $\text{gph } \varphi_{(1,2)^T, (-1,-1)^T}$ for $f(x) = 5 - \sqrt{16 - 4x_1^2 - x_2^2}$

is also convex on \mathbb{R} (i.e., an interval whenever it is not empty). Observe that, when $d = 0_n$, $I_{x,d} = \mathbb{R}$ if $x \in C$ while $I_{x,d} = \emptyset$ if $x \notin C$. Moreover, by the continuity of the function $g_{x,d}$, one has that $I_{x,d}$ is open (closed) when C is open (closed).

Definition 2.35 Let $f : C \rightarrow \mathbb{R}$, and let $x, d \in \mathbb{R}^n$ be such that $I_{x,d} \neq \emptyset$. The *section* of f corresponding to the pair x, d is defined as the function $\varphi_{x,d} : I_{x,d} \rightarrow \mathbb{R}$ given by $\varphi_{x,d} = f \circ g_{x,d}$, i.e., $\varphi_{x,d}(t) = f(x + td)$.

The graph of the section of f , $\text{gph } \varphi_{x,d}$, can be interpreted as the intersection of $\text{gph } f$ with the vertical hyperplane on the line on \mathbb{R}^n containing x and parallel to d , i.e., $\{x + td : t \in \mathbb{R}\}$; see Fig. 2.20. If $x \in C$ and $d = 0_2$, then $\varphi_{x,d}$ is the (uninteresting) function with constant value $f(x)$ on $I_{x,0_2} = \mathbb{R}$. For the function f in Fig. 2.20, if $x = (1, 2)^T$ and $d = (-1, -1)^T$, one has $I_{x,d} = \left[\frac{6}{5} - \frac{2}{5}\sqrt{19}, \frac{6}{5} + \frac{2}{5}\sqrt{19} \right]$ and $\varphi_{x,d}(t) = 5 - \sqrt{8 + 12t - 5t^2}$.

Suppose that f is differentiable at $x + td \in \text{int } C$. Then, by the chain rule, the function $\varphi_{x,d}$ is also differentiable on $t \in \text{int } I_{x,d}$ and it holds

$$\varphi'_{x,d}(t) = \nabla f(x + td)^T d = \sum_{i=1}^n \frac{\partial f(x + td)}{\partial x_i} d_i. \quad (2.20)$$

Now, suppose that f is twice differentiable at $x + td \in \text{int } C$. Then, again by the chain rule, $\varphi_{x,d}$ is always twice differentiable at $t \in \text{int } I_{x,d}$, and taking into account (2.20), it holds

$$\varphi''_{x,d}(t) = \sum_{i=1}^n \sum_{j=1}^n \frac{\partial^2 f(x + td)}{\partial x_j \partial x_i} d_i d_j = d^T \nabla^2 f(x + td) d. \quad (2.21)$$

Proposition 2.36 (Convexity of sections) *Let $C \subset \mathbb{R}^n$ be a nonempty convex set. Then the function $f : C \rightarrow \mathbb{R}$ is convex if and only if $\varphi_{x,d} : I_{x,d} \rightarrow \mathbb{R}$ is convex for all $x, d \in \mathbb{R}^n$ such that $I_{x,d} \neq \emptyset$.*

Proof Let us suppose that f is convex. Let $x, d \in \mathbb{R}^n$ be such that $I_{x,d}$ is a proper interval. Let $t, s \in I_{x,d}$ and $\mu \in]0, 1[$. Then,

$$\begin{aligned}\varphi_{x,d}((1-\mu)t + \mu s) &= f(x + ((1-\mu)t + \mu s)d) \\ &= f((1-\mu)(x + td) + \mu(x + sd)) \\ &\leq (1-\mu)f(x + td) + \mu f(x + sd) \\ &= (1-\mu)\varphi_{x,d}(t) + \mu\varphi_{x,d}(s).\end{aligned}$$

To prove the converse statement, pick any $x, y \in C$ and $\mu \in]0, 1[$. Let us consider the section $\varphi_{x,d}$, where $d = y - x$. Obviously,

$$\varphi_{x,d}(0) = f(x) \quad \text{and} \quad \varphi_{x,d}(1) = f(y). \quad (2.22)$$

As a consequence of (2.22) and of the convexity of $\varphi_{x,d}$, we have

$$\begin{aligned}f((1-\mu)x + \mu y) &= f(x + \mu d) = \varphi_{x,d}(\mu) = \varphi_{x,d}((1-\mu)0 + \mu 1) \\ &\leq (1-\mu)f(x) + \mu f(y);\end{aligned}$$

whence, f is convex. \square

The following result is an adaptation of Proposition 2.21 to functions of several variables. Definition 2.19 applies to multivariate functions just replacing $|x - a|$ by $\|x - a\|$.

Proposition 2.37 (Lipschitz continuity of convex functions) *Let C be an open convex set, and let D be a nonempty compact convex subset of C . If f is a continuously differentiable convex function on C , then f is Lipschitz continuous on D with constant*

$$L = \max\{\|\nabla f(x)\| : x \in \text{bd } D\}. \quad (2.23)$$

Proof We associate to each pair of points $x, y \in D$, $x \neq y$, the vector $d := y - x \neq 0_n$, the open interval $I_{x,d} := \{t \in \mathbb{R} : x + td \in C\}$, and the function $\varphi_{x,d} : I_{x,d} \rightarrow \mathbb{R}$ of Definition 2.35. Let $\alpha := \min\{t \in \mathbb{R} : x + td \in D\}$ and $\beta := \max\{t \in \mathbb{R} : x + td \in D\}$. By Proposition 2.21, applied to $\varphi_{x,d}$ on $[\alpha, \beta] \subset I_{x,d}$, and since $x + \alpha d, x + \beta d \in \text{bd } D$, one has

$$\begin{aligned}|f(y) - f(x)| &= |\varphi_{x,d}(1) - \varphi_{x,d}(0)| \\ &\leq \max\{|\varphi'_{x,d}(\alpha)|, |\varphi'_{x,d}(\beta)|\}|1 - 0| \\ &= \max\{|\nabla f(x + \alpha d)^T d|, |\nabla f(x + \beta d)^T d|\} \\ &\leq \|d\| \max\{\|\nabla f(x + \alpha d)\|, \|\nabla f(x + \beta d)\|\}\end{aligned}$$

$$\leq (\max\{\|\nabla f(x)\| : x \in \text{bd } D\})\|y - x\|,$$

where the existence of the last maximum is a consequence of the continuity of $\|\nabla f\|$ on the compact set $\text{bd } D$. We conclude that L in (2.23) is a Lipschitz constant for f on D . \square

Example 2.38 Let us consider $f(x) = \frac{1}{2}x^T Qx - c^T x$, where Q is an $n \times n$ symmetric positive semidefinite matrix and $c \in \mathbb{R}^n$. We shall compute Lipschitz constants for f on two compact sets of $C = \mathbb{R}^n$.

(a) If $D = \rho\mathbb{B}$, with $\rho > 0$, then

$$L = \max\{\|Qx - c\| : \|x\| = \rho\} \leq \rho\|Q\| + \|c\|. \quad (2.24)$$

In particular, if $f(x) = \|x\|^2$ and $D = \mathbb{B}$, then (2.24) provides $L = 2$. This Lipschitz constant is the best possible one because

$$\frac{|f(1, 0, \dots, 0) - f(\frac{k-1}{k}, 0, \dots, 0)|}{\|(1, 0, \dots, 0) - (\frac{k-1}{k}, 0, \dots, 0)\|} = \frac{2k-1}{k} \xrightarrow{k \rightarrow \infty} 2.$$

(b) Let $D = \text{conv}\{x_1, \dots, x_m\}$ be a polytope on \mathbb{R}^n . If $x \in D$, we can write $x = \sum_{i=1}^m \lambda_i x_i$, with $\sum_{i=1}^m \lambda_i = 1$ and $\lambda_i \geq 0$, $i = 1, \dots, m$. Then, by the convexity of the function $x \mapsto \|Qx - c\|$ on \mathbb{R}^n , one has

$$\|Qx - c\| \leq \sum_{i=1}^m \lambda_i \|Qx_i - c\| \leq \max_{i=1, \dots, m} \|Qx_i - c\|. \quad (2.25)$$

According to (2.25), one has

$$\begin{aligned} \max\{\|Qx - c\| : x \in \text{bd } D\} &\leq \max\{\|Qx - c\| : x \in D\} \\ &\leq \max_{i=1, \dots, m} \|Qx_i - c\| \\ &\leq \|c\| + \|Q\| \max_{i=1, \dots, m} \|x_i\|. \end{aligned}$$

Thus, by (2.23), the number $\|c\| + \|Q\| \max_{i=1, \dots, m} \|x_i\|$ is a Lipschitz constant for f on D . Thereby, a Lipschitz constant for $f(x) = \|x\|^2$ on $D = [-1, 1]^n$ is

$$L = \|2I_n\| \|(\pm 1, \dots, \pm 1)\| = 2\sqrt{n}.$$

The last results in this section characterize the convexity of smooth functions $f : C \rightarrow \mathbb{R}$, where C is a nonempty open convex set of \mathbb{R}^n , in terms of $\nabla f(x)$ and of $\nabla^2 f(x)$ at the points of C .

The following result contains two characterizations of the convexity of a differentiable function by means of its gradient. The first one establishes that a function is convex if and only if its graph is above the first-order approximation (i.e., the

graph is above the tangent hyperplane at every point of it). The second characterization extends the monotonicity condition of the derivative (i.e., the nondecreasing derivative) to functions of n variables.

Definition 2.39 Let $f : C \rightarrow \mathbb{R}$ be a differentiable function on C , where $\emptyset \neq C \subset \mathbb{R}^n$ is an open convex set. We say that ∇f is *monotone* on C when

$$(\nabla f(y) - \nabla f(x))^T(y - x) \geq 0, \quad \forall x, y \in C.$$

Proposition 2.40 (Characterization of differentiable convex functions) *Let $f : C \rightarrow \mathbb{R}$ be a differentiable function on C , where $\emptyset \neq C \subset \mathbb{R}^n$ is an open convex set. Then, the following statements are equivalent:*

- (i) f is convex on C .
- (ii) $f(y) \geq f(x) + \nabla f(x)^T(y - x)$ for all $x, y \in C$.
- (iii) ∇f is monotone on C .

Proof We shall prove that (i) \Rightarrow (ii) \Rightarrow (iii) \Rightarrow (i).

(i) \Rightarrow (ii) Let us suppose that f is convex on C . Let $x, y \in C$. Since C is open, $I_{x,d}$ is open too. Let us take $d := y - x$. By Proposition 2.36, we can apply Theorem 2.18(vi) to the function $\varphi_{x,d}$ at $0 \in I_{x,d}$,

$$\varphi_{x,d}(t) \geq \varphi_{x,d}(0) + \varphi'_{x,d}(0)t, \quad \forall t \in I_{x,d}.$$

Thus, since $1 \in I_{x,d}$, one has

$$\varphi_{x,d}(1) \geq \varphi_{x,d}(0) + \varphi'_{x,d}(0),$$

or, equivalently, recalling (2.20) and (2.22),

$$f(y) \geq f(x) + \nabla f(x)^T d = f(x) + \nabla f(x)^T(y - x).$$

(ii) \Rightarrow (iii) Let $x, y \in C$. Applying (ii) to the pairs x, y and y, x , we have

$$\begin{aligned} f(y) &\geq f(x) + \nabla f(x)^T(y - x) \\ &\geq f(y) + \nabla f(y)^T(x - y) + \nabla f(x)^T(y - x) \\ &= f(y) - (\nabla f(y) - \nabla f(x))^T(y - x). \end{aligned}$$

Therefore, $(\nabla f(y) - \nabla f(x))^T(y - x) \geq 0$.

(iii) \Rightarrow (i) Now, let us suppose that ∇f is monotone on C . We shall prove that all the sections of f are convex. Let $x, d \in \mathbb{R}^n$. We must prove that $\varphi'_{x,d}$ is nondecreasing on $I_{x,d}$. Let $t, s \in I_{x,d}$ be such that $t < s$. Then, $\varphi'_{x,d}(t) \leq \varphi'_{x,d}(s)$ if and only if $\nabla f(x + td)^T d \leq \nabla f(x + sd)^T d$, i.e.,

$$\left(\frac{1}{s-t} \right) (\nabla f(x+sd) - \nabla f(x+td))^T ((x+sd) - (x+td)) \geq 0,$$

which is true by the monotonicity of ∇f . \square

Remark 2.41 Observe that statement (ii) in Proposition 2.40 asserts that the first-order approximation determines a nonvertical supporting hyperplane for the epigraph at every point of the graph (see Fig. 1.12). For nondifferentiable functions, Proposition 2.33 guarantees the existence of a nonvertical supporting hyperplane at points in $\text{int } C$.

Proposition 2.42 (Characterization of twice differentiable convex functions) *Let $f : C \rightarrow \mathbb{R}$, where $C \subset \mathbb{R}^n$ is open and convex and f is twice differentiable on C . Then, f is convex on C if and only if $\nabla^2 f(x)$ is positive semidefinite for all $x \in C$.*

Proof To prove the direct implication, we suppose that $\nabla^2 f(x)$ is not positive semidefinite on C . Let $x \in C$ be such that $\nabla^2 f(x)$ is not positive semidefinite. Then, there exists $d \in \mathbb{R}^n$ such that $d^T \nabla^2 f(x)d < 0$. By (2.21), $\varphi''_{x,d}(0) = d^T \nabla^2 f(x)d < 0$, so $\varphi_{x,d}$ is not convex on $I_{x,d}$. Therefore, by Proposition 2.36, f is not convex on C either.

The converse implication can be directly proved. If $\nabla^2 f$ is positive semidefinite on C , by (2.21), for any $x, d \in \mathbb{R}^n$ it holds that $\varphi''_{x,d}(t) \geq 0$ for all $t \in I_{x,d}$. Then, by Theorem 2.24, the function $\varphi_{x,d}$ is convex, which implies that f is convex on C , again by Proposition 2.36. \square

Example 2.43 The set $C = \mathbb{R} \times \mathbb{R}_{++}$ is open and convex. The function $f : C \rightarrow \mathbb{R}$ given by $f(x, y) = \frac{x^2}{y}$ is convex on C because

$$\nabla^2 f(x, y) = \frac{2}{y^3} \begin{bmatrix} y^2 & -xy \\ -xy & x^2 \end{bmatrix}$$

is positive semidefinite on C .

Of course, if f is twice differentiable on an open and convex set C , to prove the convexity of f on C , one would normally apply Proposition 2.42. If the function is only once differentiable, it would be convenient to use Proposition 2.40, which is usually more difficult to verify in practice. If f is not even differentiable, one has no choice but to invoke Proposition 2.31 (about generation of convex functions) or Proposition 2.29 or 2.36 (which provides equivalent definitions of convex function). When C is convex but not open, the following proposition is usually used, applied to $\text{int } C$, whose closure contains C by Proposition 2.4.

Proposition 2.44 (Convexity of continuous extensions) *Let $C \subset \mathbb{R}^n$ be a nonempty convex set, and let $f : \text{cl } C \rightarrow \mathbb{R}$ be a continuous function whose restriction $f|_C$ is convex. Then, f is convex on $\text{cl } C$.*

Proof Let $x, y \in \text{cl } C$ and $\mu \in [0, 1]$. Let $\{x_k\}$ and $\{y_k\}$ be sequences in C such that $x_k \rightarrow x$ and $y_k \rightarrow y$. By the convexity of f on C , one has

$$f((1 - \mu)x_k + \mu y_k) \leq (1 - \mu)f(x_k) + \mu f(y_k), \quad \forall k \in \mathbb{N}. \quad (2.26)$$

Taking limits in (2.26) when $k \rightarrow \infty$, we obtain, by the continuity of f at $x, y \in \text{cl } C$,

$$f((1-\mu)x + \mu y) \leq (1-\mu)f(x) + \mu f(y).$$

Therefore, f is convex on $\text{cl } C$. \square

Example 2.45 The continuous function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ given by $f(x, y) = -\sqrt{|xy|}$ is not convex on \mathbb{R}^2 because

$$f\left(\frac{1}{2}(1, 1) + \frac{1}{2}(-1, 1)\right) = f(0, 1) = 0,$$

whereas

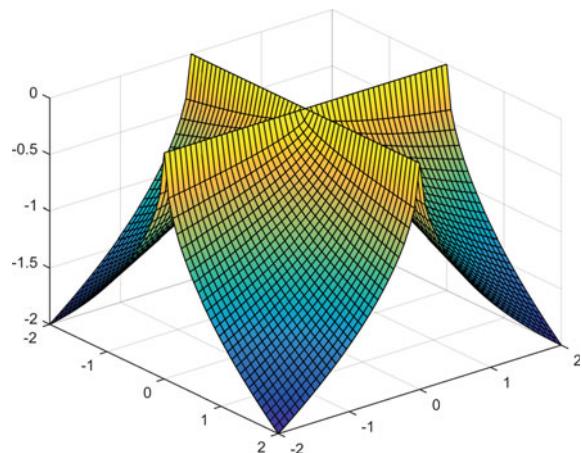
$$\frac{1}{2}f(1, 1) + \frac{1}{2}f(-1, 1) = -1.$$

The same argument proves that if f is convex on a convex set C , it must be contained in one of the four quadrants of the plane (otherwise, there exist segments contained in C whose extreme points belong to the interior of some quadrant and they intersect at least with one axis). It is easy to see that f is convex on the interior of the four quadrants because, if (x, y) belongs to one of them, then

$$\nabla^2 f(x, y) = \frac{|xy|^{-\frac{3}{2}}}{4} \begin{bmatrix} y^2 & -xy \\ -xy & x^2 \end{bmatrix},$$

which is positive semidefinite. As the quadrants are the closures of their interiors, Proposition 2.44 allows to assert that f is convex on the four quadrants (but not on their union, \mathbb{R}^2 , as shown in Fig. 2.21). Therefore, any set C such that f is convex on C must be contained in one of the four quadrants.

Fig. 2.21 Graph of $f(x, y) = -\sqrt{|xy|}$



2.4 Special Convex Functions

In this section, we introduce two classes of convex functions that are very useful to prove the existence and uniqueness of the optimal solution (which are highly desirable properties of optimization problems).

2.4.1 Strictly Convex Functions

Definition 2.46 Let $\emptyset \neq C \subset \mathbb{R}^n$ be a convex set, and let $f : C \rightarrow \mathbb{R}$. The function f is said to be *strictly convex* on C if

$$f((1 - \mu)x + \mu y) < (1 - \mu)f(x) + \mu f(y) \quad (2.27)$$

for all $\mu \in]0, 1[$ and all $x, y \in C$, with $x \neq y$.

In other words, f is strictly convex on C if f is convex on C and both sides of (2.27) coincide only when $x = y$ (i.e., $\text{gph } f$ does not contain segments).

The proofs of the results enunciated in the following examples can be obtained by changing the ones corresponding to the analogous results for convex functions.

Example 2.47 Several statements of Stolz Theorem 2.18 admit a stronger version when $f : I \rightarrow \mathbb{R}$ is strictly convex. For instance, the following statements hold:

(i) If $a, b \in I$ and $a < x < b$, then

$$\frac{f(x) - f(a)}{x - a} < \frac{f(b) - f(a)}{b - a} < \frac{f(b) - f(x)}{b - x}.$$

(ii) f'_- and f'_+ are increasing on $\text{int } I$.

(iii) If $x, y \in \text{int } I$ and $x < y$, then

$$f'_+(x) < \frac{f(y) - f(x)}{y - x} < f'_-(y).$$

Example 2.48 If I is an open interval of \mathbb{R} and $f : I \rightarrow \mathbb{R}$ is differentiable on I , then f is strictly convex on I if and only if f' is increasing on I . Moreover, f'' positive on I implies that f is strictly convex on I , but the converse statement is not true.

For example, the function $f(x) = x^4$ is strictly convex on \mathbb{R} because $f'(x) = 4x^3$ is increasing, while $f''(0) = 0$.

Example 2.49 If $\emptyset \neq C \subset \mathbb{R}^n$ is a convex set, then $f : C \rightarrow \mathbb{R}$ is strictly convex on C if and only if $\varphi_{x,d}$ is strictly convex on the interval $I_{x,d} = \{t \in \mathbb{R} : x + td \in C\}$ for any $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n \setminus \{0_n\}$ such that $I_{x,d} \neq \emptyset$.

For instance, the function $f(x) = \|x\|^2$ is strictly convex on \mathbb{R}^n because, given $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n \setminus \{0_n\}$, $\varphi_{x,d}(t) = \|x + td\|^2 = \|x\|^2 + (2d^T x)t + \|d\|^2 t^2$ satisfies $\varphi''_{x,d}(t) = 2\|d\|^2 > 0$ for all $t \in \mathbb{R}$.

Example 2.50 Let $\emptyset \neq C \subset \mathbb{R}^n$ be an open convex set, and let $f : C \rightarrow \mathbb{R}$ be twice differentiable. If $\nabla^2 f$ is positive definite for all $x \in C$, then f is strictly convex on C .

The converse statement is not true, as shown again by the function $f(x) = x^4$.

Example 2.51 Let $\emptyset \neq C \subset \mathbb{R}^n$ be an open convex set, and let $f : C \rightarrow \mathbb{R}$ be differentiable. It is easy to show that f is strictly convex on C if and only if ∇f is strictly monotone on C , which means

$$(\nabla f(y) - \nabla f(x))^T (y - x) > 0, \quad \forall x, y \in C, x \neq y.$$

If f is twice differentiable and $\nabla^2 f(x)$ is positive definite for all $x \in C$, then f is strictly convex on C . The converse does not hold.

For example, the function $f(x) = \|x\|^2$ is strictly convex on \mathbb{R}^n because, given $x, y \in \mathbb{R}^n$ such that $x \neq y$, $(\nabla f(y) - \nabla f(x))^T (y - x) = 2\|y - x\|^2 > 0$.

Example 2.52 Let $\emptyset \neq C \subset \mathbb{R}^n$ be a convex set. If $f : C \rightarrow \mathbb{R}$ is strictly convex on C , with $f(C) \subset I$ (interval in \mathbb{R}) and $h : I \rightarrow \mathbb{R}$ is convex and increasing, then $h \circ f$ is strictly convex on C . Applying this result, one obtains the strict convexity of $f : \mathbb{R}^n \rightarrow \mathbb{R}$ such that $f(x) = e^{\|x\|^2}$.

Example 2.53 Let $\emptyset \neq C \subset \mathbb{R}^n$ be a convex set, and let $f, g : C \rightarrow \mathbb{R}$ be convex functions such that at least one of them is strictly convex. Then, $f + g : C \rightarrow \mathbb{R}$ is strictly convex. By using this result, one proves that the function $f : \mathbb{R}_{++}^2 \rightarrow \mathbb{R}$ given by $f(x) = x_1^2 - 4x_1x_2 + 4x_2^2 - \ln(x_1x_2)$ is strictly convex.

The function $f : \mathbb{R}_+^2 \rightarrow \mathbb{R}$ given by $f(x, y) = -\sqrt[3]{xy}$ shows that Proposition 2.44 is not valid if one replaces “convex” by “strictly convex.” Observe also that neither the limit nor the supremum of strictly convex functions satisfies necessarily this property (consider the sequence of functions $f_k : \mathbb{R}_{++} \rightarrow \mathbb{R}$ defined by $f_k(x) = -x^{\frac{1}{k}}$, $k = 1, 2, \dots$).

2.4.2 Strongly Convex Functions

Definition 2.54 Let $\emptyset \neq C \subset \mathbb{R}^n$ be a convex set, and let $f : C \rightarrow \mathbb{R}$. The function f is said to be *strongly convex function* with coercivity coefficient $\alpha > 0$ on C if

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

for all $x, y \in C$ and $\mu \in [0, 1]$.

The strong convexity of f implies something more than strict convexity: that we can insert an arc of parabola between each arc of $\text{gph } f$ and its corresponding chord. To see it, take two arbitrary points $x, y \in C$. Let $d := y - x$. We know that $0, 1 \in I_{x,d}$. For each $t \in [0, 1]$, one has

$$\begin{aligned}\varphi_{x,d}(t) &= f((1-t)x + ty) \\ &\leq q_{x,d}(t) := (1-t)f(x) + tf(y) - \frac{\alpha}{2}t(1-t)\|x - y\|^2 \\ &\leq h_{x,d}(t) := (1-t)f(x) + tf(y),\end{aligned}$$

so that $\varphi_{x,d} \leq q_{x,d} \leq h_{x,d}$ on $[0, 1]$, where $q_{x,d}$ and $h_{x,d}$ are a convex quadratic function and an affine function on $I_{x,d}$, respectively, with $\varphi_{x,d}(0) = q_{x,d}(0) = h_{x,d}(0) = f(x)$ and $\varphi_{x,d}(1) = q_{x,d}(1) = h_{x,d}(1) = f(y)$.

Example 2.55 If $\emptyset \neq C \subset \mathbb{R}^n$ is a convex set, $f : C \rightarrow \mathbb{R}$ is strictly (strongly) convex on C and $g : C \rightarrow \mathbb{R}$ is convex on C , then $f + g$ is strictly (strongly) convex on C . Moreover, if $\emptyset \neq D \subset \mathbb{R}^m$ is a convex set and $g : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is affine and injective, with $g(D) \subset C$, then $f \circ g$ is strictly (strongly) convex on D . The injectivity condition cannot be removed, as shown by the functions $f : [1, 2] \rightarrow \mathbb{R}$ and $g : \text{conv}\{(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}), (\begin{smallmatrix} 0 \\ 1 \end{smallmatrix}), (\begin{smallmatrix} 1 \\ 1 \end{smallmatrix})\} \rightarrow \mathbb{R}$ given by $f(x) = -\ln x$ and $g(y) = y_1 + y_2$ (taking, for example, the points $(1, 0)^T, (0, 1)^T$ and $\mu = \frac{1}{2}$).

Hint: If A is a $n \times m$ full column rank matrix, then $\|Ax\| \geq \beta\|x\|$ for all $x \in \mathbb{R}^m$ where $\beta := \min\{\|Ay\| : \|y\| = 1\} > 0$.

It is easy to prove that Proposition 2.44 is true for strongly convex functions.

Proposition 2.56 (Characterization of strongly convex functions) *Let $\emptyset \neq C \subset \mathbb{R}^n$ be convex, $f : C \rightarrow \mathbb{R}$ and $\alpha > 0$. Let $g : C \rightarrow \mathbb{R}$ be such that $g := f - \frac{\alpha}{2}\|\cdot\|^2$. Then, f is strongly convex with coercivity coefficient α on C if and only if g is convex on C .*

Proof From the definition, $f = g + \frac{\alpha}{2}\|\cdot\|^2$ is strongly convex with coercivity coefficient α on C if and only if, for any $x, y \in C$ and $\mu \in [0, 1]$, it holds

$$\begin{aligned}g((1-\mu)x + \mu y) + \frac{\alpha}{2}\|(1-\mu)x + \mu y\|^2 \\ \leq (1-\mu)g(x) + \frac{\alpha}{2}(1-\mu)\|x\|^2 \\ + \mu g(y) + \frac{\alpha}{2}\mu\|y\|^2 - \frac{\alpha}{2}\mu(1-\mu)\|x - y\|^2,\end{aligned}$$

or, equivalently (reordering the terms),

$$\begin{aligned}g((1-\mu)x + \mu y) - (1-\mu)g(x) - \mu g(y) \\ \leq \frac{\alpha}{2}(1-\mu)\|x\|^2 + \frac{\alpha}{2}\mu\|y\|^2 \\ - \frac{\alpha}{2}\mu(1-\mu)\|x - y\|^2 - \frac{\alpha}{2}\|(1-\mu)x + \mu y\|^2.\end{aligned}\tag{2.28}$$

One gets the conclusion by checking the zero value of the right-hand side in (2.28), i.e., the identity

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

This completes the proof. \square

Corollary 2.57 (Characterization of twice differentiable strongly convex functions)
Let $f : C \rightarrow \mathbb{R}$, where $\emptyset \neq C \subset \mathbb{R}^n$ is open and convex and f is twice differentiable on C . Then, f is strongly convex with coercivity coefficient $\alpha > 0$ on C if and only if

$$d^T \nabla^2 f(x) d \geq \alpha \|d\|^2, \quad \forall x \in C, \forall d \in \mathbb{R}^n. \quad (2.29)$$

Proof Let $g(x) = f(x) - \frac{\alpha}{2} \|x\|^2$. Obviously, g is twice differentiable on C , with $\nabla^2 g(x) = \nabla^2 f(x) - \alpha I_n$. By Proposition 2.56, we have that the function f is strongly convex with coercivity coefficient $\alpha > 0$ on C if and only if $\nabla^2 f(x) - \alpha I_n$ is positive semidefinite for all $x \in C$, i.e., if and only if (2.29) holds. \square

Under the conditions of Corollary 2.57, according to (2.29), f is strongly convex with coercivity coefficient $\alpha > 0$ on C if and only if

$$\begin{aligned} \alpha &\leq \inf \left\{ \frac{d^T \nabla^2 f(x) d}{\|d\|^2} : x \in C, d \in \mathbb{R}^n \setminus \{0_n\} \right\} \\ &= \inf \{u^T \nabla^2 f(x) u : x \in C, u \in \mathbb{S}^{n-1}\}, \end{aligned} \quad (2.30)$$

where \mathbb{S}^{n-1} denotes the unit sphere in \mathbb{R}^n . When $n = 1$, (2.30) simplifies to

$$\alpha \leq \inf \{f''(x) : x \in C\}.$$

Corollary 2.58 If $\emptyset \neq C \subset \mathbb{R}^n$ is a bounded open and convex set and $f : C \rightarrow \mathbb{R}$ is twice differentiable on $\text{cl } C$, with $\nabla^2 f(x)$ positive definite for all $x \in \text{cl } C$, then f is strongly convex on C with coercivity coefficient

$$\min \{u^T \nabla^2 f(x) u : x \in \text{cl } C, u \in \mathbb{S}^{n-1}\} > 0.$$

Proof Let $\alpha := \inf \{u^T \nabla^2 f(x) u : x \in C, u \in \mathbb{S}^{n-1}\}$ and $Y := C \times \mathbb{S}^{n-1}$. By the assumptions, $\text{cl } Y = \text{cl } C \times \mathbb{S}^{n-1}$ is compact and the function $h : \text{cl } Y \rightarrow \mathbb{R}_+$ defined by $h(x, u) = u^T \nabla^2 f(x) u$ is continuous and positive on $\text{cl } Y$. So, $\alpha \geq \beta := \min \{h(y) : y \in \text{cl } Y\} > 0$. We have to show that $\beta \geq \alpha$.

Let $y \in \text{cl } Y$ be such that $\beta = h(y)$. Since $y \in \text{cl } Y$, there exists $\{y_k\} \subset Y$ such that $y_k \rightarrow y$. Since $\alpha \leq h(y_k)$ for all $k \in \mathbb{N}$ and $h(y_k) \rightarrow h(y)$, $\alpha \leq h(y) = \beta$. \square

Example 2.59 Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be given by $f(x_1, x_2) = (x_1 - x_2)^4 + (x_1 - 1)^4$ (the function in the Example 1.33). As $\nabla^2 f(x)$ is positive semidefinite for all $x \in \mathbb{R}^2$ whereas $\nabla^2 f(1, 1)$ is the null matrix, then (2.30) fails (i.e., $\alpha = 0$), so we have that f is not strongly convex.

Actually, $\nabla^2 f$ is not positive definite on the lines $x_1 = 1$ and $x_1 = x_2$. So, f is strongly convex on any bounded open convex set whose closure does not contain points from those lines.

Sometimes, coercivity can be ensured by computing the Hessian matrix (that is, a local property on all the points guarantees a global property).

Proposition 2.60 (Coercivity of strongly convex functions) *Let $\emptyset \neq C \subset \mathbb{R}^n$ be an unbounded convex set with nonempty interior and $f : C \rightarrow \mathbb{R}$ be continuous. If f is strongly convex on $C \subset \mathbb{R}^n$, then f is coercive.*

Proof Pick any $\bar{x} \in \text{int } C$. By Proposition 2.56, there exists $\alpha > 0$ such that the function $g(x) := f(x) - \frac{\alpha}{2} \|x\|^2$ is convex. Since f is continuous, $\text{epi } g$ is closed. Then, Proposition 2.33 implies the existence of a nonvertical hyperplane supporting $\text{cl epi } g = \text{epi } g$ at $(\bar{x}, g(\bar{x}))$. In particular, this implies the existence of $(\begin{smallmatrix} a \\ \gamma \end{smallmatrix}) \in \mathbb{R}^{n+1}$, with $\gamma > 0$, such that

$$a^T x + \gamma g(x) \geq b, \quad \forall x \in C. \quad (2.31)$$

Thus, by the definition of g together with (2.31) and the Cauchy–Schwarz inequality, we have for all $x \in C$ that

$$f(x) = \frac{\alpha}{2} \|x\|^2 + g(x) \geq \frac{\alpha}{2} \|x\|^2 + \frac{1}{\gamma} (b - a^T x) \geq \frac{\alpha}{2} \|x\|^2 - \frac{\|a\|}{\gamma} \|x\| + \frac{b}{\gamma},$$

which implies coercivity of f since $\alpha > 0$ and completes the proof. \square

Example 2.61 The exponential function is strongly convex with coercivity coefficient e^a on intervals of the form $[a, +\infty[$, but not on \mathbb{R} . As every pair of real numbers is contained in an interval of the form $[a, +\infty[$, the exponential function is strictly convex on \mathbb{R} . This example shows that strong convexity cannot be replaced by strict convexity in Proposition 2.60.

Example 2.62 The norm, $x \mapsto \|x\|$, is neither strong nor strictly convex. This example shows that the converse statement in Proposition 2.60 does not hold, even for convex functions.

2.5 Exercises

2.1 Determine whether the functions $f, g : \mathbb{R} \rightarrow \mathbb{R}$, given by $f(x) = x + \sqrt{x^2 + 4}$ and $g(x) = \ln(1 + e^x)$, are convex and Lipschitz continuous.

2.2 Consider the function $f :]-\infty, 1] \rightarrow \mathbb{R}$ given by

$$f(x) = \begin{cases} x^2 - 1, & x < 1, \\ 2, & x = 1. \end{cases}$$

Determine whether f is convex and Lipschitz continuous.

2.3 Determine whether the function $f : [1, +\infty[\rightarrow \mathbb{R}$, given by $f(x) = -\sqrt{x-1}$, is continuous, convex, strictly convex, strongly convex, and coercive.

2.4 Let $f : \mathbb{R}_{++} \rightarrow \mathbb{R}$ be such that $f(x) = x + \frac{s}{x}$.

- (a) Study whether f is convex, strictly convex, and strongly convex on \mathbb{R}_{++} .
- (b) Study whether f is coercive on \mathbb{R}_{++} .
- (c) Study whether f has a unique global minimum on \mathbb{R}_{++} .

2.5 Consider the function $f : \mathbb{R} \rightarrow \mathbb{R}$ given by $f(x) = \frac{1}{2}(\max\{x, 0\})^2$.

- (a) Determine whether f is convex, strongly convex, coercive and Lipschitz continuous.
- (b) The same as in (a) is asked in relation with its derivative f' .

2.6 Determine for which values of $p \in \mathbb{N}$ the function $f_p : I \rightarrow \mathbb{R}$, given by $f_p(x) = x^p$ for all $x \in I$, is convex, strongly convex and Lipschitz continuous at 0 when:

- (a) $I = \mathbb{R}$;
- (b) $I = [-\alpha, \alpha]$, with $\alpha > 0$.

2.7 Consider the sequence of functions $\{f_r\}$ given for $r \in \mathbb{N}$ by

$$f_r(x) = \begin{cases} \frac{1}{2}rx^2, & \text{if } |x| \leq \frac{1}{r}, \\ |x| - \frac{1}{2r}, & \text{if } |x| > \frac{1}{r}, \end{cases}$$

as well as $\lim_r f_r$, $\sup_r f_r$, and $\inf_r f_r$. Determine which of them are differentiable, continuous, and convex.

2.8 Consider the function $f : \mathbb{R}_+^2 \mapsto \mathbb{R}$ given by

$$f(x) = \begin{cases} 0, & \text{if } x_1 \geq 0 \text{ and } x_2 = 0, \\ 1, & \text{if } x_1 = 0 \text{ and } x_2 > 0, \\ 1 - \frac{x_1}{x_2}, & \text{if } x_1 > 0 \text{ and } x_2 > 0. \end{cases}$$

Determine whether f is convex and continuous on its domain.

2.9 Consider the set

$$C = \left\{ (x, y) \in]1, +\infty[^2 : (x-1)(y-1) \geq \frac{1}{4} \right\}.$$

Determine whether the function $f : C \rightarrow \mathbb{R}$, given by $f(x, y) = \frac{1}{x} + \frac{1}{y} - \frac{1}{xy}$, is convex on C .

2.10 Determine whether the function $f(x, y) = x^3 - 3xy^2$ is convex and/or coercive on \mathbb{R}^2 . Find its local and global minima on \mathbb{R}^2 .

2.11 Consider the function $f(x, y) = x^4 + y^4 - 32y^2$. Find the largest open convex sets on which f is convex and/or coercive. Compute its global minima on \mathbb{R}^2 .

2.12 Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ be such that $f(x, y) = x^2(1 + e^y)$.

(a) Find the largest open convex set $C \subset \mathbb{R}^2$ on which f is convex.

(b) Determine whether f is strongly convex and/or coercive on C .

(c) Determine whether f is Lipschitz continuous on C .

2.13 Determine whether the following function $f : \mathbb{R} \times \mathbb{R}_+ \rightarrow \mathbb{R}$ is convex:

$$f(x) = \begin{cases} \frac{x_1^2}{x_2}, & \text{if } x_2 > 0, \\ 0, & \text{if } x_2 = 0. \end{cases}$$

2.14 Consider the function $f(x) = \frac{1}{p} \sum_{i=1}^p \|x - x_i\|$ (the average of the distances to the points $x_1, \dots, x_p \in \mathbb{R}^n$).

(a) Determine whether f is continuous, convex, strictly convex, strongly convex, and coercive.

(b) Compute the global minimum set of the function f in the particular case in which $n = 1$, $p = 4$, and the points are 0, 1, 3, and 10.

(c) What is the smallest Lipschitz constant for such a function?

2.15 Consider the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$, where $f(x)$ is the greatest eigenvalue of the matrix

$$\begin{bmatrix} 0 & x_1 & 0 \\ x_1 & 0 & x_2 \\ 0 & x_2 & 0 \end{bmatrix}.$$

Determine whether f is continuous, Lipschitz continuous, differentiable, and coercive.

2.16 Let $f : \mathbb{R}_+^2 \rightarrow \mathbb{R}$ be given by

$$f(x, y) = \begin{cases} -\frac{xy}{x+y}, & \text{if } x > 0 \text{ and } y > 0, \\ 0, & \text{if } xy = 0, x \geq 0, y \geq 0. \end{cases}$$

Determine whether f is convex and strictly convex on \mathbb{R}_{++}^2 and on \mathbb{R}_+^2 .

2.17 Prove the convexity of the function $f \circ g : \mathbb{R}^2 \rightarrow \mathbb{R}$ given by

$$g(x) = x_1^2 - x_1 x_2 + x_2^2 - 2x_1 + x_2$$

and

$$f(t) = \begin{cases} 0, & \text{if } t < -\frac{\varepsilon}{q}, \\ \frac{qt^2}{2\varepsilon} + t + \frac{\varepsilon}{2q}, & \text{if } -\frac{\varepsilon}{q} \leq t \leq 0, \\ t + \frac{\varepsilon}{2q}, & \text{if } t \geq 0, \end{cases}$$

where q and ε are two given positive scalars.

2.18 Consider the function

$$f(x, y, z) = (2z - y)^2 + (x - y)^2 + (x - 1)^2.$$

- (a) Determine whether f is Lipschitz continuous on \mathbb{R}^3 .
- (b) Determine whether f is convex, strictly convex or strongly convex on \mathbb{R}^3 .
- (c) Determine whether f is coercive on \mathbb{R}^3 .

2.19 Let $f \in \mathcal{C}^2(\mathbb{R}^n)$ and $\bar{x} \in \mathbb{R}^n$. Determine whether the following statements are true or false, justifying the answer:

- (a) If $\nabla^2 f(\bar{x})$ is positive semidefinite, then f is convex on a convex neighborhood of \bar{x} .
- (b) If $\nabla^2 f(\bar{x})$ is positive definite, then f is convex on a convex neighborhood of \bar{x} .

2.20 Consider the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ given by $f(x) = \|x\|^p$, where $p \geq 2$. Analyze the validity of the following argument, in five steps, with which one aims to test the convexity of f :

- (a) $\nabla^2 f(x)$ is positive semidefinite for all $x \in \mathbb{R}^2 \setminus \{0_2\}$.
- (b) f is convex on any open convex set not containing 0_2 .
- (c) f is convex on any convex set whose interior is nonempty and does not contain 0_2 .
- (d) f is convex on the four quadrants of the plane.
- (e) f is convex on \mathbb{R}^2 .

Chapter 3

Unconstrained Optimization



This chapter studies a collection of optimization problems without functional constraints, that is, problems of the form

$$\begin{aligned} P : \text{Min } f(x) \\ \text{s.t. } x \in C, \end{aligned}$$

where $\emptyset \neq C \subset \mathbb{R}^n$ represents a given constraint set (typically \mathbb{R}^n or \mathbb{R}_{++}^n) and $f : C \rightarrow \mathbb{R}$ is the objective function. Unconstrained optimization problems have been widely used in astronomy and engineering. On the one hand, data fitting by least squares was a method developed by astronomers as Laplace and Gauss, in the second half of the eighteenth century, by using unconstrained quadratic optimization in order to get an accurate description of the behavior of celestial bodies to facilitate navigating the Earth's oceans. On the other hand, at the beginning of the twentieth century, the head of the technical department of the Danish branch of the International Bell Telephone Company, an engineer and amateur mathematician called Johan Jensen proved an inequality covering several classical ones as special cases, which allows to rigorously solve isoperimetric and design problems even though the constraint set is open. Jensen's paper "Sur les fonctions convexes et les inégalités entre les valeurs moyennes" [Acta Mathematica 30 (1906) 175–193] is generally considered the inception of convex analysis.

We shall use in this chapter three main ingredients to solve unconstrained optimization problems:

1. Analytic tools (differentiability, coercivity, and convexity).
2. Inequalities comparing different means of positive numbers.
3. Dual problems maximizing lower bounds for the objective function.

It is worth observing that one does not make an immediate use of derivatives when solving an optimization problem by means of inequalities or duality, but it is convenient to recall that most inequalities are proved by analytic arguments combin-

ing convexity and differentiability while the dual problems are frequently solved via calculus.

3.1 Unconstrained Quadratic Optimization

This section is devoted to quadratic optimization problems of the form

$$P_Q : \text{Min } f(x) = \frac{1}{2}x^T Qx - c^T x + b \\ \text{s.t. } x \in \mathbb{R}^n,$$

where $Q \in \mathcal{S}_n$, $c \in \mathbb{R}^n$, and $b \in \mathbb{R}$. The following result characterizes the relevant properties of f while the subsequent one provides closed formulas for the optimal set F^* of P_Q .

Proposition 3.1 (Convexity properties of the quadratic functions) *Let $f(x) = \frac{1}{2}x^T Qx - c^T x + b$, with $Q \in \mathcal{S}_n$, $c \in \mathbb{R}^n$, and $b \in \mathbb{R}$. Then, the following statements hold:*

- (i) *f is convex if and only if Q is positive semidefinite.*
- (ii) *f is strongly convex (strictly convex, coercive) if and only if Q is positive definite.*

Proof (i) It follows from the characterization of twice differentiable convex functions through the Hessian matrix of f (see Proposition 2.42), which is $\nabla^2 f(x) = Q$.

(ii) To prove the direct implication, if f is strongly convex or at least strictly convex, then Q is positive semidefinite by (i). We now prove by contradiction that the coercivity of f also entails that Q is positive semidefinite. Otherwise, if Q is not positive semidefinite, there exists $v \in \mathbb{R}^n \setminus \{0_n\}$ such that $v^T Qv < 0$. Then,

$$\lim_{\mu \rightarrow +\infty} f(\mu v) = \lim_{\mu \rightarrow +\infty} \left(\frac{\mu^2}{2} v^T Qv - \mu c^T v + b \right) = -\infty, \quad (3.1)$$

which contradicts the coercivity of f . Thus, we can assume that Q is positive semidefinite. We now suppose that Q is not positive definite. Since it is positive semidefinite, 0 must be the smallest eigenvalue of Q . Let $v \neq 0_n$ be a corresponding eigenvector. Then, as $Qv = 0_n$, one gets

$$f(v) = -c^T v + b = \frac{f(2v) + f(0_n)}{2},$$

so f is not even strictly convex. Consider now the vector

$$u := \begin{cases} -v, & \text{if } c^T v \leq 0, \\ v, & \text{if } c^T v > 0. \end{cases}$$

Since $\lim_{\mu \rightarrow +\infty} f(\mu u) \in \{-\infty, b\}$, with $\lim_{\mu \rightarrow +\infty} \|\mu u\| = +\infty$, f cannot be coercive.

To prove the converse implication, let U be an orthogonal $n \times n$ matrix such that

$$U^T Q U = \text{diag}(\lambda_1, \dots, \lambda_n), \quad (3.2)$$

where $\lambda_1, \dots, \lambda_n \in \mathbb{R}_{++}$ are the eigenvalues of Q (possibly repeated). Since Q is positive definite, we can assume that $0 < \lambda_{\min} := \lambda_1 \leq \dots \leq \lambda_n := \lambda_{\max}$. Recalling that $\|d\| = \|u\|$ for any pair of vectors $d, u \in \mathbb{R}^n$ such that $d = U u$, we have that (3.2) leads to

$$d^T Q d = u^T \text{diag}(\lambda_1, \dots, \lambda_n) u = \sum_{i=1}^n \lambda_i u_i^2 \geq \lambda_{\min} \|u\|^2 = \lambda_{\min} \|d\|^2 \quad (3.3)$$

for all $d \in \mathbb{R}^n$, so we have that

$$\alpha := \inf \left\{ \frac{d^T Q d}{\|d\|^2} : d \in \mathbb{R}^n \setminus \{0_n\} \right\} \geq \lambda_{\min} > 0.$$

Hence, f is strongly convex on \mathbb{R}^n with coercivity coefficient λ_{\min} . Thus, by Proposition 2.60, f is also strictly convex and coercive on \mathbb{R}^n . The proof is complete. \square

As a consequence of Proposition 3.1, the notions of coercivity, strong convexity and strict convexity coincide for quadratic functions. Denote by $Q(\mathbb{R}^n)$ the image of \mathbb{R}^n by the linear mapping $x \mapsto Qx$ and by $F^* := \arg\min\{f(x) : x \in \mathbb{R}^n\}$ the optimal set of P_Q .

Proposition 3.2 (The optimal set in quadratic optimization) *In relation with the unconstrained quadratic problem P_Q , the following statements hold:*

- (i) *If Q is not positive semidefinite, then $v(P_Q) = -\infty$.*
- (ii) *If Q is positive semidefinite, $F^* \neq \emptyset$ if and only if $c \in Q(\mathbb{R}^n)$. Then, the optimal set of P_Q is the affine manifold*

$$F^* = \{x \in \mathbb{R}^n : Qx = c\}.$$

(iii) *Q is positive definite if and only if $|F^*| = 1$. In this case,*

$$F^* = \{Q^{-1}c\}.$$

Proof (i) If Q is not positive semidefinite, there exists $v \in \mathbb{R}^n \setminus \{0_n\}$ such that $v^T Q v < 0$. Then (3.1) holds and $v(P_Q) = -\infty$.

(ii) If Q is positive semidefinite, then f is convex. By Proposition 2.40, given $\bar{x} \in \mathbb{R}^n$, one has

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}), \quad \forall x \in \mathbb{R}^n,$$

so any critical point of f is a global minimizer. Since the converse statement is always true for differentiable functions, we conclude that F^* is the set of critical points of f , i.e., $F^* = \{x \in \mathbb{R}^n : Qx = c\}$. Obviously, $F^* \neq \emptyset$ if and only if there exists $x \in \mathbb{R}^n$ such that $c = Qx$, i.e., $c \in Q(\mathbb{R}^n)$.

(iii) If Q is positive definite, then $\det Q \neq 0$ and, by (ii), we have

$$F^* = \{x \in \mathbb{R}^n : Qx = c\} = \{Q^{-1}c\},$$

so we have that $|F^*| = 1$.

Conversely, we assume that $|F^*| = 1$. This implies, by (i), that Q is positive semidefinite. Suppose that Q is not positive definite. Let $v \in \mathbb{R}^n \setminus \{0_n\}$ be an eigenvector of Q corresponding to the eigenvalue 0 and let $F^* = \{\bar{x}\}$. Then,

$$Q(\bar{x} + v) = Q\bar{x} = c,$$

so we have $\bar{x} + v \in F^*$. Thus, $|F^*| > 1$ (contradiction). \square

Example 3.3 (a) Let $f(x_1, x_2) = x_1^2 + x_2^2 + x_1x_2 + 2x_1 - x_2$. Then, Q is positive definite and

$$F^* = \left\{ Q^{-1} \begin{pmatrix} -2 \\ 1 \end{pmatrix} \right\} = \left\{ \begin{pmatrix} -\frac{5}{3} \\ \frac{4}{3} \end{pmatrix} \right\}.$$

(b) Let $f(x_1, x_2, x_3) = 2x_1^2 + x_2^2 + x_3^2 + 2x_2x_3$. Then, Q is positive semidefinite (but not positive definite) and

$$F^* = \{x \in \mathbb{R}^n : Qx = 0_n\} = \text{span}\{(0, 1, -1)^T\}.$$

3.2 Least Squares Regression

In order to detect severe acute malnutrition in infants and children, the medical societies and health departments select random samples of children whose age and weight are annotated, allowing to create growth standards for comparison purposes. Assume that the number of children in one of these samples is m , that the age is measured in months and the weight in kilograms, whose magnitudes are represented by t and s , respectively. This gives rise to a data set $\{(t_i, s_i), i = 1, \dots, m\}$ of observations, usually called point cloud, that admits a 2D plot with axis t and s . One observes that weight grows linearly with age when the range of ages is small, while this dependence would be better described by a quadratic concave function otherwise. After fitting a suitable affine or quadratic function f to the data set, the future revised children of t months will be classified as overweighted (underweighted) when its observed weight is much higher (lower) than the weight $f(t)$ predicted by the fitted model.

This same methodology is used to detect potential bank bankruptcy based on q financial ratios, which are interpreted as independent variables, while the dependent

variable could be some measure of bank's profits as the cash flow. The unique difference with the previous model is that the best fitted function f is here chosen inside certain family of functions of q variables. For the sake of simplicity, we only consider in this section models involving one independent variable.

3.2.1 Linear Regression

When the given point cloud $\{(t_i, s_i), i = 1, \dots, m\}$, with $m \geq 2$, suggests a linear dependence of s with respect to t , the regression problem consists in calculating the ordinate at the origin and the slope of the line better fitted to that point cloud, that we represent here with x_1 and x_2 as they are the decision variables in this setting. More precisely, we try to compute the pair $(x_1, x_2)^T$ minimizing some measure of the *residual vector* r , whose i -th component is $r_i = x_1 + x_2 t_i - s_i$, that is,

$$r := \begin{pmatrix} x_1 + x_2 t_1 - s_1 \\ \vdots \\ x_1 + x_2 t_m - s_m \end{pmatrix} \in \mathbb{R}^m.$$

The norms commonly used to measure r are the ℓ_∞ norm, the ℓ_1 norm, and the ℓ_2 norm.

By solving the unconstrained problem

$$P_1 : \text{Min } \|r\| = \sqrt{\sum_{i=1}^m r_i^2},$$

we get the so-called *least squares regression line*, whose objective function $\|\cdot\|$ is continuous but not differentiable on the constraint set \mathbb{R}^2 . Composing $\|\cdot\|$ with the quadratic function $y \mapsto y^2$ (increasing on \mathbb{R}_+), we obtain the equivalent unconstrained quadratic problem

$$P_2 : \text{Min } f(x_1, x_2) := \sum_{i=1}^m (x_2 t_i + x_1 - s_i)^2.$$

Denoting by $\bar{t} = \frac{1}{m} \sum_{i=1}^m t_i$ and $\bar{s} = \frac{1}{m} \sum_{i=1}^m s_i$ the observed (arithmetic) *means* of the magnitudes t and s , we can write $f(x) = \frac{1}{2} x^T Q x - c^T x + b$, with

$$Q = 2 \begin{bmatrix} m & m\bar{t} \\ m\bar{t} & \sum_{i=1}^m t_i^2 \end{bmatrix}, \quad c = 2 \begin{pmatrix} m\bar{s} \\ \sum_{i=1}^m t_i s_i \end{pmatrix} \quad \text{and} \quad b = \sum_{i=1}^m s_i^2.$$

The principal minors of Q are $2m > 0$, $2 \sum_{i=1}^m t_i^2 \geq 0$ and

$$\det Q = 4m \left(\sum_{i=1}^m t_i^2 - m\bar{t}^2 \right) = 4m \left\| (t_1 - \bar{t}, \dots, t_m - \bar{t})^T \right\|^2 \geq 0,$$

so we have that Q is positive semidefinite. Moreover, Q is positive definite if and only if $\det Q > 0$, i.e., there exist $i, j \in \{1, \dots, m\}$ such that $t_i \neq t_j$ or, equivalently, the point cloud is not vertically aligned (when the independent variable t remains fixed in all realizations of the experiment, there is no reason to infer that s depends on t , so it is statistically meaningless to look for the regression line of s on t). Consequently, f is convex; moreover, it is strongly convex and coercive except in the trivial case where the point cloud is vertically aligned. In this trivial case, $t_i = \bar{t}$ for all $i = 1, \dots, m$,

$$Q = 2m \begin{bmatrix} 1 & \bar{t} \\ \bar{t} & \bar{t}^2 \end{bmatrix}$$

and

$$c = 2 \left(\sum_{i=1}^m \frac{t_i s_i}{\bar{t}} \right) = 2\bar{s}\left(\frac{1}{\bar{t}}\right) \in \text{span}\left\{\left(\frac{1}{\bar{t}}\right)\right\} = Q(\mathbb{R}^2).$$

Therefore, P_2 has infinitely many optimal solutions, which are all couples $(x_1, x_2)^T$ such that $x_1 + \bar{t}x_2 = \bar{s}$, corresponding to nonvertical lines containing the *gravity center* $(\bar{t}, \bar{s})^T$ of the point cloud (see Fig. 3.1).

In the nontrivial case where not all t_i coincide (see Fig. 3.2), the unique optimal solution of P_2 is the solution of the so-called *normal system* $\{Qx = c\}$, that is, after

Fig. 3.1 Regression lines in the trivial case

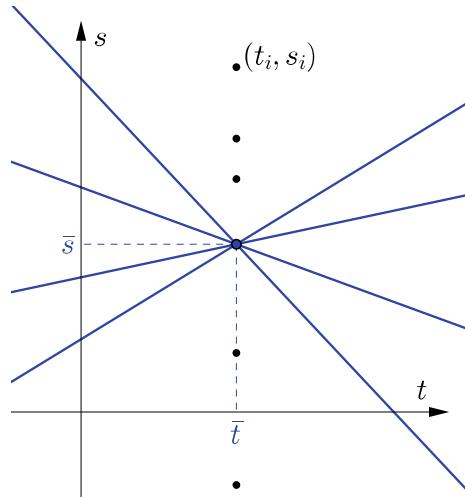
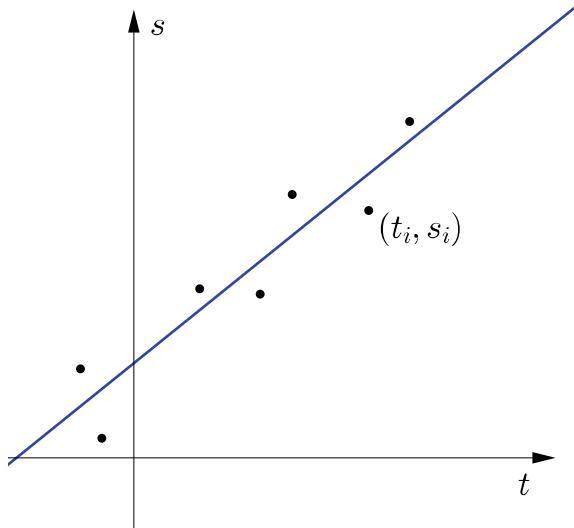


Fig. 3.2 Regression line in the nontrivial case



dividing by $2m$ both members of the first equation,

$$\begin{cases} x_1 + \bar{t}x_2 = \bar{s} \\ m\bar{t}x_1 + \left(\sum_{i=1}^m t_i^2\right)x_2 = \sum_{i=1}^m t_i s_i \end{cases},$$

whose closed solution $(\bar{x}_1, \bar{x}_2)^T$ can be easily obtained using Cramer's rule:

$$\begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix} = \begin{pmatrix} \frac{\bar{s} \sum_{i=1}^m t_i^2 - \bar{t} \sum_{i=1}^m t_i s_i}{\sum_{i=1}^m t_i^2 - m\bar{t}^2} \\ \frac{\sum_{i=1}^m t_i s_i - m\bar{t}\bar{s}}{\sum_{i=1}^m t_i^2 - m\bar{t}^2} \end{pmatrix}. \quad (3.4)$$

Linear regression is widely used in natural and social sciences, and in engineering, either directly or after a convenient transformation of the involved variables.

Example 3.4 An empirical traffic law establishes that the time needed by a vehicle to travel the stretch between two traffic lights is approximately given by

$$t = b_0 + \frac{b_1 x}{1 - \frac{x}{N}},$$

where t expresses the time in seconds, x is the number of vehicles in that stretch, and N represents the capacity of the stretch. For certain remote controlled stretch of 130 m length, with an estimated capacity of 30 vehicles, the following five observations have been recently collected.

x	20	18	22	21	15
t	14	10	16	18	8

We would like to estimate the velocity, expressed in Km/h, of certain vehicle at that stretch when there are other 11 vehicles circulating on it. After making the change of variables $y = \frac{x}{1 - \frac{x}{30}}$, equation (3.4) provides the regression line of t on y , which is $t = 2.5456 + 0.1853y$. So, the estimated time to travel along the stretch is

$$t = 2.5456 + 0.1853 \times \frac{12}{1 - \frac{12}{30}} = 6.2516 \text{ seconds},$$

with an average velocity of $\frac{130}{6.2516} \times 3.6 = 74.861 \text{ Km/h}$.

It is convenient to compare the ℓ_2 , ℓ_∞ , and ℓ_1 linear regression models. From the computational perspective, the ℓ_∞ and the ℓ_1 linear regression lines are fitted by solving the respective linear optimization problems (recall Subsection 1.1.5), while the ℓ_2 linear regression line has the closed expression $s = \bar{x}_1 + \bar{x}_2 t$, with \bar{x}_1 and \bar{x}_2 as in (3.4). To this computational advantage, we must aggregate the fact that the ℓ_2 linear regression allows to carry out statistical inference (e.g., assign confidence intervals for \bar{x}_1 and \bar{x}_2 under certain probabilistic assumptions), which explains the popularity of this type of regression among users. However, these advantages of the ℓ_2 regression have also a negative counterpart: its high sensitivity to outliers (lack of robustness), as the means square regression line is hauled by any points far away from that line, due to the power two affecting the components of the residual vector.

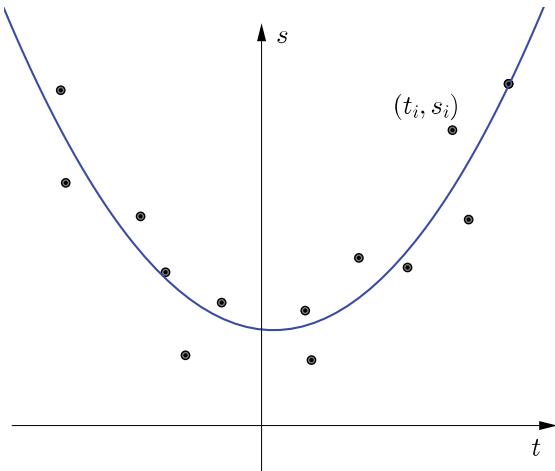
3.2.2 Polynomial Regression

Suppose now that the point cloud $\{(t_i, s_i), i = 1, \dots, m\}$ suggests a polynomial dependence of s with respect to t , of degree at most q , with $m > q \geq 1$ (in the precedent subsection, $m > q = 1$, so we are generalizing here linear regression). We are now trying to determine the polynomial $p(t) = x_1 + x_2 t + \dots + x_{q+1} t^q$ minimizing the Euclidean norm of the residual vector

$$r := \begin{pmatrix} x_1 + x_2 t_1 + \dots + x_{q+1} t_1^q - s_1 \\ \vdots \\ x_1 + x_2 t_m + \dots + x_{q+1} t_m^q - s_m \end{pmatrix} \in \mathbb{R}^m;$$

see Fig. 3.3.

Fig. 3.3 Degree 2 polynomial regression



To do this, we replace again the objective function $\|r\|$ by its square $\|r\|^2$. We will get a closed formula for the unconstrained quadratic problem

$$P_3 : \text{Min } f(x) := \sum_{i=1}^m \left(\sum_{j=1}^{q+1} x_j t_i^{j-1} - s_i \right)^2,$$

where the decision variable $x = (x_1, \dots, x_{q+1})^T$ ranges on \mathbb{R}^{q+1} . We define

$$N := \begin{bmatrix} 1 & t_1 & \dots & t_1^{q} \\ 1 & t_2 & \dots & t_2^{q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & \dots & t_m^{q} \end{bmatrix} \quad \text{and} \quad s := \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{pmatrix}$$

(observe that the determinants of the submatrices of N of order $q + 1$ are of Vandermonde's type). It can be easily observed that

$$f(x) = \|Nx - s\|^2 = \|Nx\|^2 - 2s^T Nx + \|s\|^2 = x^T (N^T N)x - 2(N^T s)^T x + \|s\|^2,$$

so we can write $f(x) = \frac{1}{2}x^T Qx - c^T x + b$, with $Q = 2N^T N$, $c = 2N^T s$, and $b = \|s\|^2$. By Proposition 1.19, since $\frac{1}{2}Q$ is the Gram matrix of N^T , Q is positive semidefinite and f is convex; moreover, Q is positive definite if and only if the columns of N are linearly independent, which holds if and only if $\{|t_1, \dots, t_m|\} \geq q + 1$ (i.e., the projection of the point cloud onto the t -axis space contains at least $q + 1$ different points). When $q = 1$ (linear regression), this assumption means that the point

cloud is not vertically aligned (recall the discussion on the trivial case in the last subsection).

Suppose that $|\{t_1, \dots, t_m\}| \geq q + 1$. As Q is positive definite, by Proposition 3.2, the unique optimal solution of P_3 is the solution of the *normal system*

$$\{(N^T N)x = N^T s\}, \quad (3.5)$$

that is,

$$\bar{x} = (N^T N)^{-1} N^T s. \quad (3.6)$$

We thus have the closed formula (3.6) for the computation of the polynomial of degree at most q , $p(t) = \bar{x}_1 + \bar{x}_2 t + \dots + \bar{x}_{q+1} t^q$, which provides a best least squares approximation to the point cloud $\{(t_i, s_i), i = 1, \dots, m\}$. In practice, in order to avoid the high computational cost of inverting the matrix $(N^T N)^{-1}$, it is preferable to solve (3.5) instead of using the closed formula (3.6). Notice also that we recover (3.4) whenever $q = 1$.

If $|\{t_1, \dots, t_m\}| < q + 1$, then there exist infinitely many ℓ_2 regression polynomials of degree at most q : those of the form $p(t) = \bar{x}_1 + \bar{x}_2 t + \dots + \bar{x}_{q+1} t^q$, with $(\bar{x}_1, \dots, \bar{x}_{q+1})^T$ satisfying the consistent indeterminate linear system (3.5) $\{(N^T N)x = N^T s\}$.

Example 3.5 Suppose we are required to fit a cubic polynomial to certain 2D point cloud of the form $\{(i - 3, s_i), i = 1, \dots, 5\}$, where $\{s_1, \dots, s_5\} \subset \mathbb{R}$ are to be obtained experimentally. The best least squares approximation to this parameterized cloud is the polynomial $p(t) = \bar{x}_1 + \bar{x}_2 t + \bar{x}_3 t^2 + \bar{x}_4 t^3$, where $\bar{x} = (N^T N)^{-1} N^T s$, with $s^T = (s_1, \dots, s_5)$,

$$N = \begin{bmatrix} 1 & -2 & 4 & -8 \\ 1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \end{bmatrix} \quad \text{and} \quad (N^T N)^{-1} = \begin{bmatrix} \frac{17}{35} & 0 & -\frac{1}{7} & 0 \\ 0 & \frac{65}{72} & 0 & -\frac{17}{72} \\ -\frac{1}{7} & 0 & \frac{1}{14} & 0 \\ 0 & -\frac{17}{72} & 0 & \frac{5}{72} \end{bmatrix}.$$

Thus,

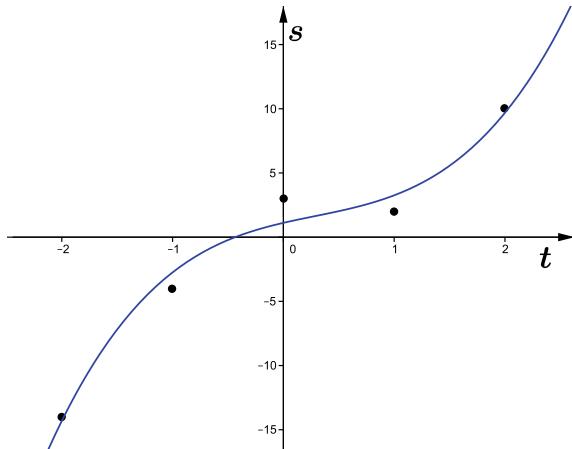
$$\bar{x} = \begin{bmatrix} -\frac{3}{35} & \frac{12}{35} & \frac{17}{35} & \frac{12}{35} & -\frac{3}{35} \\ \frac{1}{12} & -\frac{2}{3} & 0 & \frac{2}{3} & -\frac{1}{12} \\ \frac{1}{7} & -\frac{1}{14} & -\frac{1}{7} & -\frac{1}{14} & \frac{1}{7} \\ -\frac{1}{12} & \frac{1}{6} & 0 & -\frac{1}{6} & \frac{1}{12} \end{bmatrix} s$$

i.e., the best degree 3 polynomial approximation to the point cloud is

$$p(t) = \left(-\frac{3}{35}, \frac{12}{35}, \frac{17}{35}, \frac{12}{35}, -\frac{3}{35}\right)s + \left(\frac{1}{12}, -\frac{2}{3}, 0, \frac{2}{3}, -\frac{1}{12}\right)st + \left(\frac{1}{7}, -\frac{1}{14}, -\frac{1}{7}, -\frac{1}{14}, \frac{1}{7}\right)st^2 + \left(-\frac{1}{12}, \frac{1}{6}, 0, -\frac{1}{6}, \frac{1}{12}\right)st^3.$$

For instance, the ℓ_2 cubic regression for the point cloud

Fig. 3.4 ℓ_2 cubic regression of a point cloud



$$\{(-2, -14), (-1, -4), (0, 3), (1, 2), (2, 10)\}$$

is $p(t) = t^3 - \frac{6}{7}t^2 + 2t + \frac{39}{35}$ (see Figure 3.4).

3.3 Least Squares Solutions to Overdetermined Linear Systems

The location of wild animals in natural parks is usually made by means of vehicles equipped with a global positioning system (GPS) and a directional antenna that receives greater power in the direction of the microtransmitter previously installed at the animal's collars. Then, each received signal determines a line that is assumed to pass close to the animal's location. The equations of these lines provide an overdetermined system of $m > 2$ equations and two unknowns (the coordinates of the animal's position) to be approximately solved.

The same methodology was used by Gauss in 1801 to predict the location of the dwarf planet (asteroid) Ceres, with the unique difference that the lines were traced with telescopes and the number of unknowns was three. To approximately solve an overdetermined system $\{Ax = b\}$, one has to minimize some norm of the residual vector $r := Ax - b$. Laplace considered the Euclidean and the ℓ_∞ norms. In the latter case, the optimization problem can be reformulated as a linear one (recall Subsection 1.1.5).

In this section, following Gauss' approach, we consider the ℓ_2 norm, as the problem has then a closed solution when the columns of A are linearly independent, even though this solution is less robust than the one provided by the ℓ_∞ norm. We associate with the matrix A ($m \times n$) and the vector $b \in \mathbb{R}^m$ the unconstrained optimization problem

$$\begin{aligned} P_1 : \text{Min } & \|Ax - b\| \\ \text{s.t. } & x \in \mathbb{R}^n. \end{aligned}$$

Proposition 3.6 (Approximate solution of inconsistent systems) *The optimal solutions of P_1 are the solutions of the consistent linear system*

$$\{(A^T A)x = A^T b\}. \quad (3.7)$$

If the columns of A are linearly independent, then the unique optimal solution of P_1 is

$$\bar{x} := (A^T A)^{-1} A^T b. \quad (3.8)$$

Proof We replace P_1 by the equivalent unconstrained quadratic optimization problem

$$\begin{aligned} P_2 : \text{Min } & f(x) = \|Ax - b\|^2 \\ \text{s.t. } & x \in \mathbb{R}^n. \end{aligned}$$

We can write $f(x) = x^T A^T Ax - 2(A^T b)^T x + \|b\|^2$, where $A^T A$ is positive semidefinite (and positive definite whenever the columns of A are linearly independent). The conclusion follows from Proposition 3.2, taking into account the well-known identity $(A^T A)(\mathbb{R}^n) = A^T(\mathbb{R}^n)$. \square

As usual in least squares methods, it is preferable in practice to compute \bar{x} as the solution of the system in (3.7) instead of using the closed formula (3.8).

Example 3.7 Consider the inconsistent system $\{x + y = 2, x + 2y = 3, x + 3y = 3\}$. We have

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad b = \begin{pmatrix} 2 \\ 3 \\ 3 \end{pmatrix}, \quad A^T A = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}, \quad \text{and } A^T b = \begin{pmatrix} 8 \\ 17 \end{pmatrix}.$$

By solving the system (3.7), or using the closed formula (3.8) with

$$(A^T A)^{-1} A^T b = \begin{pmatrix} \frac{4}{3} & \frac{1}{3} & -\frac{2}{3} \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 3 \end{pmatrix},$$

one gets the best least squares solution of the above system: $(\frac{5}{3}, \frac{1}{2})^T$ (see Fig. 3.5).

It is widely recognized that *machine learning* and *big data* applications give rise to challenging optimization problems [18]. For instance, text classification leads to convex optimization problems of high dimensionality. Most of these problems come from the ubiquitous linear observation model

$$b = Ax + z, \quad (3.9)$$

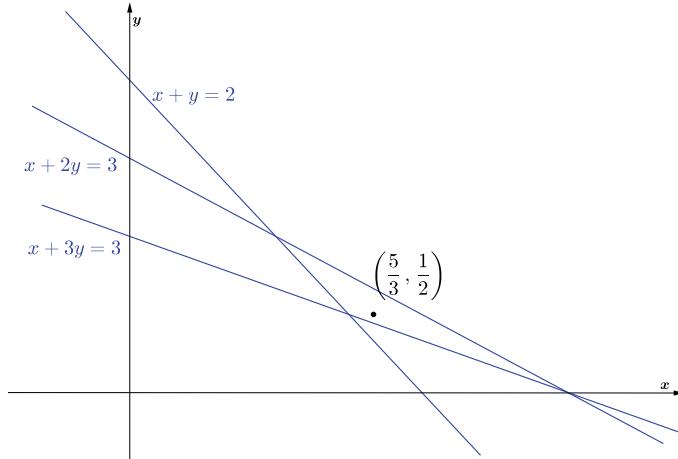


Fig. 3.5 Best least squares solution of an inconsistent system

where $b \in \mathbb{R}^m$ is the vector of observations, A is a known $m \times n$ matrix, x is an unknown parameter vector, and $z \in \mathbb{R}^m$ is a vector of unknown perturbations or noise.

The linear model (3.9), along with low-dimensionality requirements on x , such as sparsity, low total-variation, low-rankness, etc., have been areas of intense research activity in signal processing. The classical convex formulation in this setting has always been the least squares approach, leading to the unconstrained quadratic optimization problem

$$P_{LS} : \text{Min}_{x \in \mathbb{R}^n} f(x) = \frac{1}{2} \|Ax - b\|^2. \quad (3.10)$$

For large dimensions, the unconstrained convex quadratic problems can be efficiently solved by Krylov subspace methods (see, e.g., [88, 89]).

3.4 Optimizing Without Derivatives

This section is devoted to solve some isoperimetric and geometric optimization problems via inequalities and duality.

3.4.1 Jensen's Inequalities

We prove in this subsection the discrete form of a famous inequality proved by Jensen in 1906 together with its continuous version (relating the value of a convex function with an integral of the function).

Proposition 3.8 (1st Jensen inequality) Let $\emptyset \neq C \subset \mathbb{R}^n$ be a convex set and let $f : C \rightarrow \mathbb{R}$. Then, f is convex on C if and only if for any finite set of points $x_1, \dots, x_m \in C$ and corresponding nonnegative scalars μ_1, \dots, μ_m such that $\sum_{i=1}^m \mu_i = 1$, the following inequality holds:

$$f\left(\sum_{i=1}^m \mu_i x_i\right) \leq \sum_{i=1}^m \mu_i f(x_i). \quad (3.11)$$

Moreover, if f is strictly convex, (3.11) holds with equality if and only $x_1 = \dots = x_m$.

Proof To demonstrate the characterization, we only need to show the direct statement, and we shall prove it by induction on the number m of elements of C . It is true for convex combinations of pairs of elements of C by Definition 2.26 of convexity. Assume then that it is true for all convex combinations of at most m elements of C .

Suppose first that f is convex. Let any $x_1, \dots, x_{m+1} \in C$ and consider some scalars $\mu_1, \dots, \mu_{m+1} \in \mathbb{R}_+$ such that $\sum_{i=1}^{m+1} \mu_i = 1$. We can assume that $\mu_1, \dots, \mu_{m+1} \in \mathbb{R}_{++}$ (otherwise we can apply the induction assumption). Consider $\gamma = \sum_{i=2}^{m+1} \mu_i \in]0, 1[$. By the induction assumption, applied twice, we have

$$\begin{aligned} f\left(\sum_{i=1}^{m+1} \mu_i x_i\right) &= f\left(\mu_1 x_1 + \gamma \sum_{i=2}^{m+1} \frac{\mu_i}{\gamma} x_i\right) \leq \mu_1 f(x_1) + \gamma f\left(\sum_{i=2}^{m+1} \frac{\mu_i}{\gamma} x_i\right) \\ &\leq \mu_1 f(x_1) + \gamma \sum_{i=2}^{m+1} \frac{\mu_i}{\gamma} f(x_i) = \sum_{i=1}^{m+1} \mu_i f(x_i). \end{aligned} \quad (3.12)$$

Assume now that f is strictly convex and that the second statement holds for convex combinations of at most m elements of C . Suppose that

$$f\left(\sum_{i=1}^{m+1} \mu_i x_i\right) = \sum_{i=1}^{m+1} \mu_i f(x_i). \quad (3.13)$$

From (3.13), all inequalities in (3.12) (which holds due to the convexity of f) are equalities, that is,

$$f\left(\mu_1 x_1 + \gamma \sum_{i=2}^{m+1} \frac{\mu_i}{\gamma} x_i\right) = \mu_1 f(x_1) + \gamma f\left(\sum_{i=2}^{m+1} \frac{\mu_i}{\gamma} x_i\right)$$

and

$$f\left(\sum_{i=2}^{m+1} \frac{\mu_i}{\gamma} x_i\right) = \sum_{i=2}^{m+1} \frac{\mu_i}{\gamma} f(x_i),$$

whence, again by the induction assumption, $x_1 = \sum_{i=2}^{m+1} \frac{\mu_i}{\gamma} x_i = x_2 = \dots = x_{m+1}$. The proof is complete. \square

Corollary 3.9 (Arithmetic-quadratic inequality) *Given some positive numbers x_1, \dots, x_n and some corresponding scalars $\mu_1, \dots, \mu_n \in \mathbb{R}_+$ with $\sum_{i=1}^n \mu_i = 1$, it holds*

$$\sum_{i=1}^n \mu_i x_i \leq \sqrt{\sum_{i=1}^n \mu_i x_i^2},$$

and the equality is satisfied if and only if $x_1 = \dots = x_n$.

Proof Taking $f(x) = x^2$ in the 1st Jensen inequality, one gets

$$\left(\sum_{i=1}^n \mu_i x_i \right)^2 = f \left(\sum_{i=1}^n \mu_i x_i \right) \leq \sum_{i=1}^n \mu_i f(x_i) = \sum_{i=1}^n \mu_i x_i^2, \quad (3.14)$$

and the equality holds if and only if $x_1 = \dots = x_n$. We get the aimed conclusion by taking square roots at both sides of (3.14). \square

Example 3.10 Let $h : \mathbb{R}^3 \rightarrow \mathbb{R}$ be given by $h(x) = \frac{x_1^2}{a^2} + \frac{x_2^2}{b^2} + \frac{x_3^2}{c^2}$, where a, b, c are some positive scalars. The point of the ellipsoid $\{x \in \mathbb{R}^3 : h(x) \leq 1\}$ maximizing

$$f(x) := \nabla h(x)^T \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = 2 \left(\frac{x_1}{a^2} + \frac{x_2}{b^2} + \frac{x_3}{c^2} \right)$$

belongs to $\{x \in \mathbb{R}_{++}^3 : h(x) = 1\}$. For any x in the latter set, if $k := 2 \left(\frac{1}{a^2} + \frac{1}{b^2} + \frac{1}{c^2} \right)$, the arithmetic-quadratic inequality yields

$$f(x) = k \left(\frac{2x_1}{ka^2} + \frac{2x_2}{kb^2} + \frac{2x_3}{kc^2} \right) \leq k \sqrt{\frac{2x_1^2}{ka^2} + \frac{2x_2^2}{kb^2} + \frac{2x_3^2}{kc^2}} = \sqrt{2kh(x)} = \sqrt{2k},$$

and the equality is satisfied when $x_1 = x_2 = x_3$. Therefore, the maximum of f on $\{x \in \mathbb{R}^3 : h(x) \leq 1\}$ is attained at x such that $x_1 = x_2 = x_3 = \sqrt{\frac{2}{k}}$.

Corollary 3.11 (2nd Jensen inequality) *Given some positive numbers x_1, \dots, x_n and some corresponding scalars $\mu_1, \dots, \mu_n \in \mathbb{R}_{++}$ such that $\sum_{i=1}^n \mu_i = 1$, it holds*

$$\prod_{i=1}^n x_i \leq \sum_{i=1}^n \mu_i x_i^{\frac{1}{\mu_i}},$$

and the equality is satisfied if and only if $x_1 = \dots = x_n$.

Proof By applying the 1st Jensen inequality to the exponential function, which is strictly convex, we get

$$\begin{aligned} \prod_{i=1}^n x_i &= \exp\left(\ln\left(\prod_{i=1}^n x_i\right)\right) = \exp\left(\sum_{i=1}^n \ln x_i\right) = \exp\left(\sum_{i=1}^n \mu_i \ln x_i^{\frac{1}{\mu_i}}\right) \\ &\leq \sum_{i=1}^n \mu_i \exp\left(\ln x_i^{\frac{1}{\mu_i}}\right) = \sum_{i=1}^n \mu_i x_i^{\frac{1}{\mu_i}}, \end{aligned}$$

and the equality holds if and only if $x_1 = \dots = x_n$. \square

Corollary 3.12 (Geometric-quadratic inequality) *Given some positive numbers x_1, \dots, x_n and some corresponding scalars $\mu_1, \dots, \mu_n \in \mathbb{R}_{++}$ such that $\sum_{i=1}^n \mu_i = 1$, it holds*

$$\prod_{i=1}^n x_i^{\mu_i} \leq \sqrt{\sum_{i=1}^n \mu_i x_i^2},$$

and the equality is satisfied if and only if $x_1 = \dots = x_n$.

Proof Let $y_i := x_i^{\mu_i} > 0$, $i = 1, \dots, n$. By the 2nd Jensen and the arithmetic-quadratic inequalities,

$$\prod_{i=1}^n x_i^{\mu_i} = \prod_{i=1}^n y_i \leq \sum_{i=1}^n \mu_i y_i^{\frac{1}{\mu_i}} = \sum_{i=1}^n \mu_i x_i \leq \sqrt{\sum_{i=1}^n \mu_i x_i^2},$$

and the equality holds if and only if $x_1 = \dots = x_n$. \square

Example 3.13 Consider the isoperimetric problem consisting in maximizing the volume of the cuboids (rectangular boxes) inscribed in a sphere given its radius. When $n = 3$, we can assume without loss of generality that the sphere is centered at 0_3 and has radius 1 and that the vertices of the inscribed cuboids are expressed as $(\pm x_1, \pm x_2, \pm x_3)^T$, with $x_1, x_2, x_3 \in \mathbb{R}_{++}$ and $\|x\| = 1$, with a volume of $8x_1 x_2 x_3$ units. By the geometric-quadratic inequality, with $n = 3$ and $\mu_1 = \mu_2 = \mu_3 = \frac{1}{3}$, one has

$$\sqrt[3]{x_1 x_2 x_3} \leq \sqrt{\frac{1}{3}(x_1^2 + x_2^2 + x_3^2)} = \frac{\sqrt{3}}{3},$$

that is, $8x_1 x_2 x_3 \leq \frac{8\sqrt{3}}{9}$, and the maximum of the first member is attained when $x_1 = x_2 = x_3 = \frac{\sqrt{3}}{3}$ (i.e., when the cuboid is a cube). This means in practice (sculpture, gemology) that, to get the greatest cuboid block from a spheric solid of radius R , one has to carve a cube of edges $\frac{2\sqrt{3}}{3}R$, which only occupies the 36.8% of the initial solid's volume.

3.4.2 The Geometric-Arithmetic Inequality

The most widely used inequality in optimization generalizes the well-known inequality $\sqrt{x_1 x_2} \leq \frac{x_1 + x_2}{2}$ between the geometric and the arithmetic means of two positive numbers x_1 and x_2 .

Proposition 3.14 (Geometric-arithmetic inequality) *Given some positive numbers x_1, \dots, x_n and some corresponding scalars $\mu_1, \dots, \mu_n \in \mathbb{R}_{++}$ such that $\sum_{i=1}^n \mu_i = 1$, it holds*

$$\prod_{i=1}^n x_i^{\mu_i} \leq \sum_{i=1}^n \mu_i x_i, \quad (3.15)$$

and the equality is satisfied if and only if $x_1 = \dots = x_n$.

Proof The function $f(x) = -\ln x$ is strictly convex on \mathbb{R}_{++} as $f''(x) = \frac{1}{x^2} > 0$ for all $x \in \mathbb{R}_{++}$. Let $x_1, \dots, x_n \in \mathbb{R}_{++}$ and $\mu_1, \dots, \mu_n \in \mathbb{R}_{++}$ be such that $\sum_{i=1}^n \mu_i = 1$. By Proposition 3.8 (1st Jensen inequality), one has

$$-\ln\left(\sum_{i=1}^n \mu_i x_i\right) = f\left(\sum_{i=1}^n \mu_i x_i\right) \leq \sum_{i=1}^n \mu_i f(x_i) = -\sum_{i=1}^n \mu_i \ln x_i, \quad (3.16)$$

and the equality holds if and only if $x_1 = \dots = x_n$. The inequality (3.16) is equivalent to

$$\ln\left(\sum_{i=1}^n \mu_i x_i\right) \geq \sum_{i=1}^n \mu_i \ln x_i = \ln \prod_{i=1}^n x_i^{\mu_i}. \quad (3.17)$$

Finally, since the exponential function is increasing, we deduce that (3.17) is equivalent to (3.15). \square

Obviously, the geometric-quadratic inequality is a straightforward consequence of the geometric-arithmetic and the arithmetic-quadratic inequalities.

The geometric-arithmetic inequality (3.15) not only allows to solve simple isoperimetric problems, but also design problems arising in engineering for which the calculus rules may provide critical points, but do not guarantee optimality and uniqueness. When dealing with minimization (maximization) problems, the strategy consists in rewriting the objective function (a posynomial) as a positive multiple of a weighted arithmetic (geometric) mean in such a way that the corresponding geometric (arithmetic, respectively) mean is constant. As an example, let us revisit the design problem of Example 1.6, which was initially formulated as

$$\begin{aligned} P_1 : \text{Min } f(x) &= x_1 x_2 + x_1 x_3 + x_2 x_3 \\ \text{s.t. } h(x) &= x_1 x_2 x_3 = 1, \\ x &\in \mathbb{R}_{++}^3. \end{aligned}$$

We now reformulate it as

$$\begin{aligned} P_2 : \text{Min } f(x) &= 3\left(\frac{1}{3}x_1x_2 + \frac{1}{3}x_1x_3 + \frac{1}{3}x_2x_3\right) \\ \text{s.t. } x &\in C, \end{aligned}$$

where $C := \{x \in \mathbb{R}_{++}^3 : x_1x_2x_3 = 1\}$. Then, by the geometric-arithmetic inequality,

$$f(x) = 3\left(\frac{1}{3}x_1x_2 + \frac{1}{3}x_1x_3 + \frac{1}{3}x_2x_3\right) \geq 3(x_1x_2)^{\frac{1}{3}}(x_1x_3)^{\frac{1}{3}}(x_2x_3)^{\frac{1}{3}} = 3(x_1x_2x_3)^{\frac{2}{3}} = 3,$$

for all $x \in C$, so the minimum of f is attained for those $x \in C$ such that $x_1x_2 = x_1x_3 = x_2x_3$, that is, at $\bar{x} = (1, 1, 1)^T$, which is the unique global minimum of P_1 .

The same method allows to solve the geometric optimization problem posed in Exercise 1.4, namely,

$$\begin{aligned} P : \text{Min } f(x) &= \frac{40}{x_1x_2x_3} + x_1x_2 + 4x_1x_3 + 4x_2x_3 \\ \text{s.t. } x &\in \mathbb{R}_{++}^3, \end{aligned}$$

where x_1 , x_2 , and x_3 represent the edges' length of the rectangular box expressed in meters. Observing that the product of the summands in the above expression of f is not constant, but becomes constant by replacing the first summand, $\frac{40}{x_1x_2x_3}$, by the sum $\frac{20}{x_1x_2x_3} + \frac{20}{x_1x_2x_3}$, we get the inequality

$$\begin{aligned} f(x) &= 5\left(\frac{1}{5}\frac{20}{x_1x_2x_3} + \frac{1}{5}\frac{20}{x_1x_2x_3} + \frac{1}{5}(x_1x_2) + \frac{1}{5}(4x_1x_3) + \frac{1}{5}(4x_2x_3)\right) \\ &\geq 5(2^2 5)^{\frac{2}{5}} (2^2)^{\frac{2}{5}}, \end{aligned}$$

so we have that the unique optimal solution of P is the solution of the system of nonlinear equations

$$\frac{20}{x_1x_2x_3} = x_1x_2 = 4x_1x_3 = 4x_2x_3,$$

that is,

$$\bar{x} = \left(2^{\frac{4}{5}} 5^{\frac{1}{5}}, 2^{\frac{4}{5}} 5^{\frac{1}{5}}, 2^{-\frac{6}{5}} 5^{\frac{1}{5}}\right)^T \simeq (2.4022, 2.4022, 0.6006)^T.$$

Example 3.15 We try to determine the triangle of a given perimeter with the greatest area. We can assume the perimeter has length 2. If the side lengths are x, y, z , the area of such a triangle is, by the famous Heron's formula, $\frac{1}{2}\sqrt{(1-x)(1-y)(1-z)}$, where $1-x = \frac{x+y+z}{2} - x = \frac{y+z-x}{2}$ is positive (as any side length is less than the sum of the two other side lengths), as well as $1-y$ and $1-z$. Defining $C := \{(x, y, z) \in [0, 1]^3 : x + y + z = 2\}$, we have to solve the problem

$$\begin{aligned} P_1 : \text{Max } f(x, y, z) &= \frac{1}{2}\sqrt{(1-x)(1-y)(1-z)} \\ \text{s.t. } (x, y, z) &\in C, \end{aligned}$$

or the equivalent one

$$\begin{aligned} P_2 : \text{Max } f(x, y, z) &= \sqrt[3]{(1-x)(1-y)(1-z)} \\ \text{s.t. } (x, y, z) &\in C. \end{aligned}$$

By the geometric-arithmetic inequality, applied to the numbers $1-x$, $1-y$, and $1-z$, with equal weights $\frac{1}{3}$, one gets

$$\sqrt[3]{(1-x)(1-y)(1-z)} \leq \frac{(1-x) + (1-y) + (1-z)}{3} = \frac{1}{3}$$

for all $(x, y, z) \in C$, and the equality holds whenever $1-x = 1-y = 1-z$, that is, when $x = y = z = \frac{2}{3}$ (for an equilateral triangle).

The main difficulty with the described method is the selection of the suitable weights for the objective functions providing a constant value at the other member of the geometric-arithmetic inequality. These weights are obtained in a systematic way in the next section.

3.4.3 Duality in Unconstrained Geometric Optimization*

We start by recalling the concept of posynomial introduced in Subsection 1.1.4.

Definition 3.16 A *monomial* of n variables x_1, \dots, x_n is an expression of the form

$$c \prod_{j=1}^n x_j^{\alpha_j},$$

where $c \in \mathbb{R}$ is called the *coefficient* while $\alpha_1, \dots, \alpha_n \in \mathbb{R}$ are the exponents. A *posynomial* is the sum of finitely many monomials with positive coefficients.

Any monomial defines a real-valued function on \mathbb{R}_{++}^n (but maybe not on the whole of \mathbb{R}^n). Hence, any posynomial defines a function $f : \mathbb{R}_{++}^n \rightarrow \mathbb{R}$ that can be expressed as

$$f(x) = \sum_{i=1}^m c_i \prod_{j=1}^n x_j^{\alpha_{ij}},$$

with coefficients $c_i \in \mathbb{R}_{++}$ and exponents $\alpha_{ij} \in \mathbb{R}$, $i = 1, \dots, m$, $j = 1, \dots, n$.

We consider in this section problems of the form

$$\begin{aligned} P_G : \text{Min } f(x) &= \sum_{i=1}^m c_i \prod_{j=1}^n x_j^{\alpha_{ij}} \\ \text{s.t. } x &\in \mathbb{R}_{++}^n, \end{aligned}$$

which can be solved in an algebraic way by means of the geometric-arithmetic inequality.

Observe that $f(x) > 0$ for all $x \in \mathbb{R}_{++}^n$, so one has $v(P_G) \geq 0$. However, the existence of an optimal solution depends on the data, e.g., if $n = 1$ and $f(x_1) = x_1$ ($f(x_1) = x_1 + (x_1)^{-1}$), the optimal set is $F^* = \emptyset$ ($F^* = \{1\}$, respectively). Let us now discuss the viability of solving P_G by appealing to analytic tools. The differentiability of f is of course guaranteed, but not its convexity or coercivity. In fact, the natural way to prove the convexity or coercivity of f would be to prove the same properties of the monomials $u_i(x) := c_i \prod_{j=1}^n x_j^{\alpha_{ij}}$, $i = 1, \dots, m$, which would imply the convexity and coercivity of their sum. The function $h_i(x) := \ln u_i(x) = \sum_{j=1}^n \alpha_{ij} \ln x_j + \ln c_i$ is convex (concave) if $\alpha_{ij} \leq 0$ for all j ($\alpha_{ij} \geq 0$ for all j , respectively), in which case $u_i(x) = e^{h_i(x)}$ is convex (concave) too. But, unfortunately, the exponents in the monomials have in general different signs and f is usually neither convex nor concave. In the same way, the functions u_i are not coercive whenever some of their exponents are negative, so we have that f is, in general, non-coercive as well. We could also explore the reformulation of P_G through the change of variables $x_j = e^{y_j}$, $j = 1, \dots, n$, which provides the equivalent unconstrained convex optimization problem (by Proposition 2.31),

$$\begin{aligned}\tilde{P}_G : \text{Min } f(y) &= \sum_{i=1}^m c_i \exp\left(\sum_{j=1}^n \alpha_{ij} y_j\right) \\ \text{s.t. } y &\in \mathbb{R}^n,\end{aligned}$$

which is more convenient, even though computing the critical points of \tilde{P}_G may be a hard task (one has to use the available software for systems of nonlinear equations). Hence, instead of appealing to differentiability, convexity and coercivity arguments, we shall handle P_G via duality, i.e., by computing a best lower bound for f .

We now associate with P_G a suitable dual problem as follows. Given m real numbers $y_1 > 0, \dots, y_m > 0$ (*positivity condition*), called dual variables, we can write

$$f(x) = \sum_{i=1}^m y_i \left(\frac{c_i \prod_{j=1}^n x_j^{\alpha_{ij}}}{y_i} \right).$$

By aggregating the *normality* condition $\sum_{i=1}^m y_i = 1$, we can apply the geometric-arithmetic inequality (3.15) as follows:

$$\begin{aligned}f(x) &\geq \prod_{i=1}^m \left(\frac{c_i \prod_{j=1}^n x_j^{\alpha_{ij}}}{y_i} \right)^{y_i} = \prod_{i=1}^m \left(\frac{c_i}{y_i} \right)^{y_i} \left(\prod_{i=1}^m \prod_{j=1}^n x_j^{\alpha_{ij} y_i} \right) \\ &= \prod_{i=1}^m \left(\frac{c_i}{y_i} \right)^{y_i} \left(\prod_{j=1}^n (x_j)^{\sum_{i=1}^m \alpha_{ij} y_i} \right).\end{aligned}$$

By imposing the *orthogonality* condition $\sum_{i=1}^m \alpha_{ij} y_i = 0$, $j = 1, \dots, n$, one has

$$f(x) \geq \prod_{i=1}^m \left(\frac{c_i}{y_i} \right)^{y_i}. \quad (3.18)$$

We already have the lower bounds for f required to define a (*geometric*) *dual problem* D_G for P_G . Its objective function will be $g(y) := \prod_{i=1}^m \left(\frac{c_i}{y_i} \right)^{y_i}$, and the constraints will consist in the positivity, normality, and orthogonality conditions:

$$\begin{aligned} D_G : \text{Max } g(y) &= \prod_{i=1}^m \left(\frac{c_i}{y_i} \right)^{y_i} \\ \text{s.t. } &\sum_{i=1}^m y_i = 1, \\ &\sum_{i=1}^m \alpha_{ij} y_i = 0, j = 1, \dots, n, \\ &y \in \mathbb{R}_{++}^m. \end{aligned}$$

In practice, to solve D_G analytically, one reduces the dimension of the problem by exploiting the linearity of the normality and orthogonality conditions. We denote by G the feasible set of D_G , by G^* its optimal set and by $v(D_G)$ its optimal value. From (3.18), we get the *weak duality theorem*:

$$v(D_G) \leq v(P_G).$$

By defining $u_i(x) := c_i \prod_{j=1}^n x_j^{\alpha_{ij}}$, $i = 1, \dots, m$, we can write $f(x) = \sum_{i=1}^m u_i(x)$.

Theorem 3.17 (Strong duality) *If $F^* \neq \emptyset$, then $G \neq \emptyset$. Moreover, if $\bar{x} \in F^*$,*

$$\bar{y}^T := \left(\frac{u_1(\bar{x})}{f(\bar{x})}, \dots, \frac{u_m(\bar{x})}{f(\bar{x})} \right) \in G^*,$$

with $g(\bar{y}) = f(\bar{x})$, so one has that $G^* \neq \emptyset$ and $v(D_G) = v(P_G)$.

Proof The assumption that $\bar{x} \in F^*$ implies that $\bar{x} \in \mathbb{R}_{++}^n$ and $u_i(\bar{x}) \in \mathbb{R}_{++}$, $i = 1, \dots, m$, so $\bar{y} \in \mathbb{R}_{++}^m$.

Observe that, given $i \in \{1, \dots, m\}$, $j \in \{1, \dots, n\}$, one has

$$\frac{\partial u_i}{\partial x_j} = c_i \alpha_{ij} x_j^{\alpha_{ij}-1} \prod_{\substack{k=1, \dots, n \\ k \neq j}} x_k^{\alpha_{ik}},$$

so we have that

$$x_j \frac{\partial u_i}{\partial x_j} = \alpha_{ij} u_i. \quad (3.19)$$

Moreover, since $\bar{x} \in F^*$ and the constraint set $C := \mathbb{R}_{++}^n$ is open, \bar{x} must be a critical point for f . Then, by (3.19), we have

$$0 = \frac{\partial f(\bar{x})}{\partial x_j} = \sum_{i=1}^m \frac{\partial u_i(\bar{x})}{\partial x_j} = \frac{1}{\bar{x}_j} \sum_{i=1}^m \alpha_{ij} u_i(\bar{x}),$$

which entails

$$\sum_{i=1}^m \alpha_{ij} u_i(\bar{x}) = 0, \quad \forall j = 1, \dots, n. \quad (3.20)$$

Dividing by $f(\bar{x}) > 0$ both members of (3.20), we get

$$\sum_{i=1}^m \alpha_{ij} \bar{y}_i = 0, \quad \forall j = 1, \dots, n,$$

so the vector \bar{y} satisfies the orthogonality condition together with the positivity. Moreover,

$$\sum_{i=1}^m \bar{y}_i = \sum_{i=1}^m \frac{u_i(\bar{x})}{f(\bar{x})} = 1$$

(normality), which implies $\bar{y} \in G$.

It remains to prove that $f(\bar{x}) = g(\bar{y})$. Observe that the orthogonality of \bar{y} leads to

$$\prod_{i=1}^m \left(\frac{u_i(\bar{x})}{c_i} \right)^{\bar{y}_i} = \prod_{i=1}^m \prod_{j=1}^n \bar{x}_j^{\alpha_{ij} \bar{y}_i} = \prod_{j=1}^n \prod_{i=1}^m \bar{x}_j^{\alpha_{ij} \bar{y}_i} = \prod_{j=1}^n \bar{x}_j^0 = 1,$$

so we have that

$$\prod_{i=1}^m u_i(\bar{x})^{\bar{y}_i} = \prod_{i=1}^m c_i^{\bar{y}_i}. \quad (3.21)$$

From the normality of \bar{y} , the definition of this vector and (3.21), one gets

$$f(\bar{x}) = \prod_{i=1}^m f(\bar{x})^{\bar{y}_i} = \prod_{i=1}^m \left(\frac{u_i(\bar{x})}{\bar{y}_i} \right)^{\bar{y}_i} = \prod_{i=1}^m \left(\frac{c_i}{\bar{y}_i} \right)^{\bar{y}_i} = g(\bar{y}),$$

which completes the proof. \square

This strong duality theorem can be used to solve P_G as follows. If $G = \emptyset$, then $F^* = \emptyset$ (end). Otherwise, if $G \neq \emptyset$, compute an optimal solution \bar{y} of D_G (unnecessary whenever G is a singleton set) and find then the aimed optimal solution of P_G by solving the system of linear equations that results from applying logarithms to the nonlinear equations

$$u_i(x) = g(\bar{y})\bar{y}_i, \quad i = 1, \dots, m. \quad (3.22)$$

We illustrate this method with the reformulation

$$\begin{aligned} P : \text{Min } f(x) &= x_1 x_2 + \frac{1}{x_1} + \frac{1}{x_2} \\ \text{s.t. } x &\in \mathbb{R}_{++}^2 \end{aligned}$$

of the design problem in Example 1.6. Since the dual feasible set G is formed by the solutions of the linear system

$$\begin{cases} y_1 + y_2 + y_3 = 1 \\ y_1 - y_2 = 0 \\ y_1 - y_3 = 0 \end{cases},$$

we have $G = \left\{ \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right)^T \right\}$, with $g\left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3} \right) = 3$. Then (3.22) reduces to

$$x_1 x_2 = \frac{1}{x_1} = \frac{1}{x_2} = 1,$$

whose unique solution is $x_1 = x_2 = 1$ (we do not need to take logarithms in this case). Hence, $F^* = \{(1, 1, 1)^T\}$.

In a similar way, the reader can verify that the unique optimal solution of the dual problem corresponding to the geometric optimization problem in Exercise 1.4 is $\bar{y}^T = \left(\frac{2}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5} \right)$.

3.5 Exercises

3.1 Analyze the convexity and the strict and strong convexity, as well as the coercivity of the following functions on \mathbb{R}^2 , and find their global minimum sets.

- (a) $f(x_1, x_2) = 5x_1^2 + 2x_1 x_2 + x_2^2 - x_1 + 2x_2 + 3$.
- (b) $f(x_1, x_2) = \frac{1}{2}x_1^2 + \frac{3}{2}x_2^2 + \sqrt{3}x_1 x_2$.
- (c) $f(x_1, x_2) = (x_1 - x_2)^2 + (x_1 + 2x_2 + 1)^2 - 8x_1 x_2$.

3.2 Obtain the ℓ_2 linear regression line for the following point cloud:

$$\{(-1, 2), (0, 1), (1, 0), (7, 8)\}.$$

3.3 Determine the quadratic function being the best ℓ_2 approximation for the following point cloud:

$$\{(-2, 5), (-1, -1), (0, 4), (1, 7), (2, 6), (3, 5), (4, -1)\}.$$

3.4 Determine the polynomial of degree at most 4 which is the best least squares approximation of the point cloud

$$\{(-2, -2), (-1, -1), (0, 1), (1, 1), (2, 2)\}.$$

3.5 Compute the best least squares solution of the following inconsistent systems:

(a) $\{x_1 + x_2 = 1; x_1 + x_2 = -1\}$.

(b)

$$\left\{ \begin{array}{rcl} x_1 & +x_2 & +x_3 = 3 \\ & & x_3 = 1 \\ x_1 & & +x_3 = 2 \\ 2x_1 & & +5x_3 = 8 \\ -7x_1 & +8x_2 & = 0 \\ x_1 & +x_2 & -x_3 = 1 \end{array} \right\}.$$

3.6 Find the best ℓ_2 linear approximation to the function $f(t) = \frac{1}{t^2+1}$ on the interval $[0, 1]$, i.e., the affine function $g(t) = x_1 t + x_2$ (depending on the decision variables x_1 and x_2) that minimizes the norm of the residue

$$\|g - f\|_2 = \sqrt{\int_0^1 |g(t) - f(t)|^2 dt}.$$

3.7 Find the point in \mathbb{R}^n minimizing the sum of the square distances to m given points x^1, \dots, x^m . Interpret geometrically the optimal solution when $m = 4$ and these points form a quadrilateral or a parallelogram.

3.8 Prove that, from all the rectangles with a given perimeter, the one with the largest area is the square, by using derivatives, by completing the square, and by means of the geometric-arithmetic inequality.

3.9 Let A , B , and C be the vertices of a given acute triangle. Find the point P in the interior of the triangle that maximizes the product of the distances to the sides. Help: express the area of the triangle as the sum of the areas of the triangle with vertices P and each couple of $\{A, B, C\}$.

3.10 Solve the following geometric optimization problem by means of the geometric-arithmetic inequality:

$$\begin{aligned} P_G : \text{Min } f(x) &= 3x^3 + \frac{5}{x} \\ \text{s.t. } x &\in \mathbb{R}_{++}. \end{aligned}$$

3.11 Solve the following geometric optimization problem by means of the geometric-arithmetic inequality:

$$\begin{aligned} P_G : \text{Min } f(x) &= 4x_1 + \frac{x_1}{x_2^2} + \frac{4x_2}{x_1} \\ \text{s.t. } x &\in \mathbb{R}_{++}^2. \end{aligned}$$

3.12 Solve the problem

$$\begin{aligned} P_G : \text{Max } & xyz \\ \text{s.t. } & 3x + 4y + 12z = 1, \\ & x > 0, y > 0, z > 0. \end{aligned}$$

3.13 Prove, by using the geometric-arithmetic inequality, that if we have two closed cans with the same volume and height, one of them with circular base and the other with rectangular base, the one with circular base always has lower lateral area than the one with rectangular base.

3.14 Solve the geometric optimization problem:

$$\begin{aligned} P_G : \text{Min } & f(x) = \frac{50}{x_1} + \frac{20}{x_2} + x_1 x_2 \\ \text{s.t. } & x \in \mathbb{R}_{++}^2. \end{aligned}$$

3.15 Design an open topped cuboid box with a given surface area $S > 0$ having maximum volume. Also, design a similar box with a given capacity $V > 0$ having minimum surface area.

3.16 Design the biggest cylindrical tank you can build with a fixed cost of $c_0 \in \mathbb{E}$, knowing that the cost per surface unit of the top and bottom faces is $c_1 \in \mathbb{E}$ and the one of the lateral surface is $c_2 \in \mathbb{E}$. Also, design the cheapest tank that can be built with such materials having a given volume V_0 .

3.17 Prove that, given $u, v \in \mathbb{R}^n$ and $p, q \in [1, +\infty[$ such that $\frac{1}{p} + \frac{1}{q} = 1$, it holds

$$\sum_{i=1}^n |u_i v_i| \leq \|u\|_p \|v\|_q \quad (\text{Hölder inequality}),$$

where

$$\|u\|_p := \left(\sum_{i=1}^n |u_i|^p \right)^{\frac{1}{p}} \quad \text{and} \quad \|v\|_q := \left(\sum_{i=1}^n |v_i|^q \right)^{\frac{1}{q}}$$

are the ℓ_p and ℓ_q norms of u and v , respectively.

3.18 Prove that, given $x, y \in \mathbb{R}^n$ and $p \in \mathbb{N}$, it holds $\|x + y\|_p \leq \|x\|_p + \|y\|_p$ (triangle inequality property of the ℓ_p norm, also called Minkowski inequality).

3.19 Prove the Cauchy–Schwarz inequality, according to which, $|x^T y| \leq \|x\| \|y\|$ for all pair $x, y \in \mathbb{R}^n$ ([19, Subsections 1.8.2 and 1.8.3] describes applications of the Cauchy–Schwarz inequality to statistical estimation and signal processing).

3.20 Solve the geometric optimization problem

$$\begin{aligned} P_G : \text{Min } & f(x) = \frac{1000}{x_1 x_2} + 2x_1 + 2x_2 + x_1 x_2 \\ \text{s.t. } & x \in \mathbb{R}_{++}^2. \end{aligned}$$

3.21 Solve the problem

$$\begin{aligned} P_G : \text{Min } f(x) &= \frac{2}{x_1 x_2} + x_1 x_2 + x_1 \\ \text{s.t. } x &\in \mathbb{R}_{++}^2, \end{aligned}$$

by using its dual problem.

3.22 Solve the geometric optimization problem

$$\begin{aligned} P_G : \text{Min } f(x) &= c_1 x_1 + c_2 x_1^{-2} x_2^{-3} + c_3 x_2^4 \\ \text{s.t. } x &\in \mathbb{R}_{++}^2, \end{aligned}$$

where c_1 , c_2 , and c_3 are some given positive constants.

3.23 Solve the geometric optimization problem

$$\begin{aligned} P_G : \text{Min } f(x) &= c_1 x_3^{-1} + c_2 x_2^{-6} x_4^2 + c_3 x_1^3 x_4^2 + c_4 x_1^{-1} x_2^4 x_3^2 x_4^2 \\ \text{s.t. } x &\in \mathbb{R}_{++}^4, \end{aligned}$$

where c_1 , c_2 , c_3 , and c_4 are some given positive constants.

Chapter 4

Convex Optimization



Convex optimization deals with problems of the form

$$\begin{aligned} P : \text{Min } & f(x) \\ \text{s.t. } & x \in F, \end{aligned} \tag{4.1}$$

where $\emptyset \neq F \subset \mathbb{R}^n$ is a convex set and $f : F \rightarrow \mathbb{R}$ is a convex function. In this chapter, we analyze four particular cases of problem P in (4.1):

- (a) *Unconstrained convex optimization*, where the constraint set F represents a given convex subset of \mathbb{R}^n (as \mathbb{R}_{++}^n).
- (b) *Convex optimization with linear constraints*, where

$$F = \{x \in \mathbb{R}^n : g_i(x) \leq 0, i \in I\},$$

with $I = \{1, \dots, m\}$, $m \geq 1$, and g_i are affine functions for all $i \in I$. In this case, F is a polyhedral convex set (an affine manifold in the particular case where F is the solution set of a system of linear equations, as each equation can be replaced by two inequalities).

- (c) *Convex optimization with inequality constraints*, where

$$F = \{x \in C : g_i(x) \leq 0, i \in I\},$$

with $\emptyset \neq C \subset \mathbb{R}^n$ being a given constraint convex set ($C = \mathbb{R}^n$ by default) and $g_i : C \rightarrow \mathbb{R}$ are convex functions for all $i \in I$. Observe that $F = \cap_{i \in I} S_0(g_i)$ is convex as each sublevel set $S_0(g_i) = \{x \in C : g_i(x) \leq 0\}$ is convex by the convexity of g_i , $i \in I$. If C is closed and g_i is continuous on C for all $i \in I$, then F is closed by the closedness of $S_0(g_i)$, $i \in I$ (if $C = \mathbb{R}$, all constraint functions g_i are continuous due to their Lipschitz continuity on open bounded intervals; in fact, it can be proved that any convex function on $C = \mathbb{R}^n$ is also continuous).

Sensitivity analysis gives an answer to “what if” questions as “how does small perturbations in some part of the data affect the optimal value of P ?”.

tackles this question when the perturbations affect the right-hand side of the constraints (representing the available resources when P models a production planning problem). Subsections 4.4.3 and 4.4.4 introduce two different dual problems for P and provide their corresponding strong duality theorems.

(d) *Conic optimization*, where F is defined by means of constraints involving closed convex cones in certain Euclidean spaces. This type of convex optimization problems is very important in practice and can be solved by means of efficient numerical methods based on advanced mathematical tools. Since they can hardly be solved with pen and paper, the purpose of Section 4.5 is to introduce the readers in this fascinating field through the construction of duality problems for this class of problems and the review of the corresponding strong duality theorems.

4.1 The Optimal Set

The next result concerns the optimal set $F^* = \operatorname{argmin}\{f(x) : x \in F\}$ of P .

Proposition 4.1 (The optimal set in convex optimization) *Let P be the convex optimization problem in (4.1). Then:*

- (i) *F^* is a convex set.*
- (ii) *If f is strictly convex on F , then $|F^*| \leq 1$.*
- (iii) *If F is an unbounded closed set such that $\operatorname{int} F \neq \emptyset$ and f is strongly convex and continuous on F , then $|F^*| = 1$.*
- (iv) *If f is differentiable on $\operatorname{int} F \neq \emptyset$ and continuous on F , then*

$$\{x \in \operatorname{int} F : \nabla f(x) = 0_n\} \subset F^*.$$

- (v) *If F is open and f is differentiable on F , then*

$$F^* = \{x \in F : \nabla f(x) = 0_n\}.$$

Proof (i) We can assume that $F^* \neq \emptyset$. Then, due to the convexity of F and f , $F^* = S_{v(P)}(f)$ is convex.

(ii) Assume that f is strictly convex on F and there exist $x, y \in F^*$, with $x \neq y$. Then, $\frac{x+y}{2} \in F^*$ and

$$f\left(\frac{x+y}{2}\right) < \frac{1}{2}f(x) + \frac{1}{2}f(y) = v(P),$$

which is a contradiction.

(iii) Since F is an unbounded closed set and f is coercive by Proposition 2.60, we have $|F^*| \geq 1$. Then, $|F^*| = 1$ by (ii).

(iv) We must show that any critical point $\bar{x} \in \text{int } F$ of f is a global minimum. In fact, by Proposition 2.40, for any $x \in \text{int } F$, one has

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x}) = f(\bar{x}). \quad (4.2)$$

Given $x \in F$, as $F \subset \text{cl } F = \text{cl int } F$ by Proposition 2.4, there exists a sequence $\{x_r\} \subset \text{int } F$ such that $x_r \rightarrow x$. By (4.2), $f(\bar{x}) \leq f(x_r)$ for all r and, taking limits, we get $f(\bar{x}) \leq f(x)$. Hence, $\bar{x} \in F^*$.

(v) The inclusion $\{x \in F : \nabla f(x) = 0_n\} \subset F^*$ is consequence of (iv) applied to all $x \in F$, while the reverse inclusion follows from Fermat's principle. \square

Example 4.2 In Example 1.40, we considered the function

$$f(x) = \sqrt{x^2 + a^2} + \sqrt{(x - 1)^2 + b^2},$$

with $a, b \in \mathbb{R}$, $a < b$, in order to prove the reflection law. Since f'' is positive on \mathbb{R} , f is strictly convex on \mathbb{R} . Thus, f attains its minimum at a unique point (the critical point $\bar{x} = \frac{a}{a+b}$).

The same happens with the function $f(x) = \frac{\sqrt{x^2+a^2}}{v_1} + \frac{\sqrt{(x-1)^2+b^2}}{v_2}$ used to prove the refraction law.

Example 4.3 Let $f : F = \mathbb{R} \times]0, 2[\rightarrow \mathbb{R}$ be given by $f(x, y) = \ln y + \frac{1+x^2}{y}$. Since

$$\nabla^2 f(x, y) = \begin{bmatrix} \frac{2}{y} & -\frac{2x}{y^2} \\ -\frac{2x}{y^2} & \frac{2(x^2+1)}{y^3} - \frac{1}{y^2} \end{bmatrix} \Rightarrow \det \nabla^2 f(x, y) = \frac{2}{y^3} \left(\frac{2}{y} - 1 \right) > 0$$

for all $y \in]0, 2[$, $\nabla^2 f$ is positive definite on F and f is strictly convex on F . According to Proposition 4.1(ii), $|F^*| \leq 1$. Indeed, by (v), $F^* = \{(0, 1)^T\}$. Observe that the existence of optimal solution does not follow from (iii) as F is not closed and f is not strongly convex because

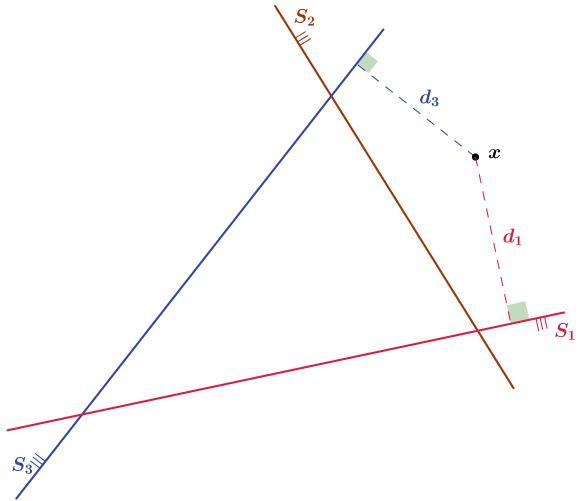
$$\lim_{r \rightarrow \infty} \nabla^2 f\left(0, 2 - \frac{1}{r}\right) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

It is easy to see that the inclusion in Proposition 4.1(iv) may be strict (consider $f(x) = e^x$ with F being any closed interval in \mathbb{R}).

We now show an immediate application of this result to systems of linear inequalities, while the next two sections provide applications to statistical inference and operations research.

A problem arising in signal processing consists in the search of points in the intersection of finitely many closed half-spaces which, frequently, do not have common points. This problem can be formulated as the computation of an approximate solution (in the least squares sense) to a (possibly inconsistent) linear system

Fig. 4.1 Illustration of $(Ax - b)_+$



$\{Ax \leq b\}$, where A is an $m \times n$ matrix, $b \in \mathbb{R}^m$, and \leq denotes the partial ordering in \mathbb{R}^m such that $y \leq z$ when $y_i \leq z_i$ for all $i = 1, \dots, m$. In formal terms,

$$P_1 : \text{Min}_{x \in \mathbb{R}^n} \| (Ax - b)_+ \|,$$

where

$$(Ax - b)_+ := (\max\{a_1^T x - b_1, 0\}, \dots, \max\{a_m^T x - b_m, 0\})^T,$$

with a_i^T denoting the i th row of A , $i = 1, \dots, m$.

Let us illustrate the meaning of $(Ax - b)_+$ with a simple example, with $n = 2$, $m = 3$ and $\|a_i\| = 1$, $i = 1, 2, 3$. Let $S_i := \{x \in \mathbb{R}^2 : a_i^T x \leq b_i\}$, $i = 1, 2, 3$, and assume that $\cap_{i=1}^3 S_i = \emptyset$. Given $x \in \mathbb{R}^2$, we have (see, e.g., (4.19) below) that

$$\max\{a_i^T x - b_i, 0\} = \begin{cases} a_i^T x - b_i = d(x, \text{bd } S_i), & \text{if } x \notin S_i, \\ 0, & \text{otherwise.} \end{cases}$$

Hence, for x as in Fig. 4.1, we get $(Ax - b)_+ = (d_1, 0, d_3)^T$.

Corollary 4.4 (Least squares solution to an inconsistent linear inequality system)
The optimal solutions to P_1 are the solutions to the nonlinear system of equations

$$A^T (Ax - b)_+ = 0_n. \quad (4.3)$$

Proof We replace P_1 by the equivalent unconstrained optimization problem

$$P_2 : \text{Min}_{x \in \mathbb{R}^n} f(x) = \| (Ax - b)_+ \|^2.$$

Denoting by $p_+ : \mathbb{R} \rightarrow \mathbb{R}$ the *positive part function*, that is, $p_+(y) := \max\{y, 0\}$ and by $h_i(x) : \mathbb{R}^n \rightarrow \mathbb{R}$ the affine function $h_i(x) := a_i^T x - b_i$, $i = 1, \dots, m$, we can write $f = \sum_{i=1}^m (p_+^2 \circ h_i)$. Since p_+^2 is convex and differentiable, with

$$\frac{dp_+^2(y)}{dy} = 2p_+(y), \quad \forall y \in \mathbb{R},$$

while $\nabla h_i(x) = a_i$ for all $x \in \mathbb{R}^n$, $i = 1, \dots, m$, we deduce that f is convex and differentiable as it is the sum of m functions satisfying these properties, with

$$\nabla f(x) = 2 \sum_{i=1}^m (a_i^T x - b_i)_+ a_i = 2A^T(Ax - b)_+.$$

The conclusion follows from Proposition 4.1(v). \square

Three elementary proofs of the existence of solution for the nonlinear system in (4.3) can be found in [25], but the analytical computation of such a solution is usually a hard task. The uniqueness is not guaranteed as the objective function of P_2 is not strictly convex.

Example 4.5 The optimality condition (4.3) for the inconsistent system

$$\{x_1 \leq -1, -x_1 \leq -1, -x_2 \leq -1, x_2 \leq -1\}$$

reduces to

$$\begin{pmatrix} (x_1 + 1)_+ - (1 - x_1)_+ \\ -(1 - x_2)_+ + (x_2 + 1)_+ \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

i.e., $(x_1 + 1)_+ = (1 - x_1)_+$ and $(x_2 + 1)_+ = (1 - x_2)_+$, whose unique solution is $(0, 0)$ (see Fig. 4.2).

Example 4.6 Consider now the inconsistent system

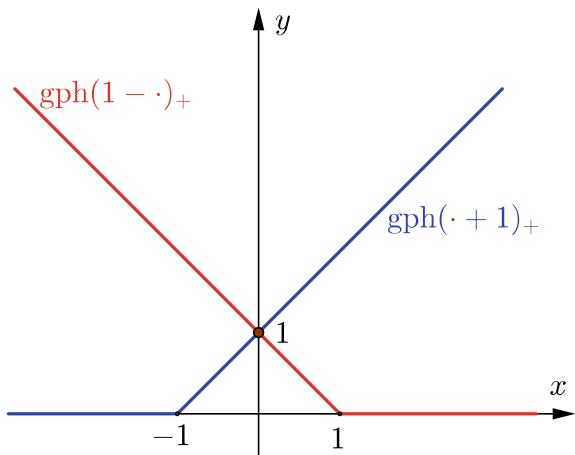
$$\{x_1 \leq -1, -x_1 \leq -1, x_2 \leq 1\}.$$

The optimality condition (4.3) is

$$\begin{pmatrix} (x_1 + 1)_+ - (-x_1 + 1)_+ \\ (x_2 - 1)_+ \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

which holds if and only if $x_1 = 0$ and $x_2 \leq 1$. Then, the set of least squares solutions of the above system is $\{0\} \times]-\infty, 1]$.

Fig. 4.2 Solution of
 $(x + 1)_+ = (1 - x)_+$



4.2 Unconstrained Convex Optimization

This section presents two interesting applications of unconstrained convex optimization. The first one is a classical location problem posed by Fermat in the seventeenth century that gave rise to a vast amount of the literature illustrating the three paradigms of optimization: the geometrical, the analytical, and the numerical ones. The second application consists in a detailed resolution of an important optimization problem arising in statistical inference that it is commonly solved in a nonrigorous way.

4.2.1 The Fermat–Steiner Problem

The following problem was first posed by Fermat to Torricelli: Given three points in the plane P_1 , P_2 , and P_3 , compute a fourth one P that minimizes the sum of distances to the first three points, i.e.,

$$P_{FS} : \text{Min } f(P) := d(P, P_1) + d(P, P_2) + d(P, P_3),$$

where $d(P, P_i)$ denotes the distance from P to P_i , $i = 1, 2, 3$. This type of problem frequently arises in practical situations, where P may represent the point where a water well should be sunk, a fire station must be installed, a rural hospital should be constructed, etc., with the purpose of serving three close villages. We can assume that P_1 , P_2 , and P_3 are not aligned (otherwise, the optimal solution of P_{FS} is obviously the point, of the three given points, which belongs to the segment determined by the two other points).

The problem P_{FS} appears under different names in the mathematical literature (Fermat, Fermat–Steiner, Fermat–Torricelli, Fermat–Steiner–Weber, etc.), and its

variants, generically called location (or p -center) problems, are still being studied by operational researchers and analysts. The most prominent of these variants are those that consider more than three given points, those which assign weights to the given points (e.g., number of residents at each village in the examples above), those which replace the Euclidean distance with non-Euclidean ones, those which replace the plane by spaces of greater (even infinite) dimension, those which replace the objective function by another one as $d(P, P_1)^2 + d(P, P_2)^2 + d(P, P_3)^2$ (in which case it is easy to prove that the optimal solution is the barycenter of the triangle of vertices P_1 , P_2 , and P_3) or $\max\{d(P, P_1), d(P, P_2), d(P, P_3)\}$ (frequently used when locating emergency services), etc. There exists an abundant literature on the characterization of the triangle centers as optimal solutions to suitable unconstrained optimization problems. One of the last centers to be characterized in this way has been the orthocenter (see [55]).

Solving an optimization problem by means of the geometric approach starts with a preliminary empirical phase, which is seldom mentioned in the literature. This provides a conjecture on the optimal solution (a point in the case of P_{FS}), which is followed by the conception of a rigorous compass and ruler constructive proof, in the style of Euclid's elements. The limitation of this approach is the lack of a general method to get conjectures and proofs, which obliges the decision maker to conceive ad hoc experiments and proofs. We illustrate the geometric approach by solving P_{FS} for triples of points P_1 , P_2 , and P_3 determining obtuse-angled triangles. We denote by α_i the angle (expressed in degrees) corresponding to vertex P_i , $i = 1, 2, 3$. Intuition suggests that when $\max\{\alpha_1, \alpha_2, \alpha_3\}$ is big enough, the optimal solution is the vertex corresponding to the obtuse angle. Indeed, we now show that this conjecture is true when $\max\{\alpha_1, \alpha_2, \alpha_3\} \geq 120$ degrees by means of a compass and ruler constructive proof.

Proposition 4.7 (Trivial Fermat–Steiner problems) *If $\alpha_i \geq 120$ degrees, then P_i is a global minimum of P_{FS} .*

Proof Assume that the angle at P_2 , say α_2 , is not less than 120 degrees. By applying to P_1 and to an arbitrary point $P \neq P_2$ a counterclockwise rotation centered at P_2 , with angle $\alpha := 180 - \alpha_2$ degrees, one gets the image points P'_1 and P' , so that the points P'_1 , P_2 , and P_3 are aligned (see Fig. 4.3).

We now recall two known facts: First, the rotations preserve the Euclidean distance between pairs of points, and second, in any isosceles triangle whose unequal angle measures $0 < \alpha \leq 60$ degrees, the length of the opposite side is less or equal than the length of the equal sides. Hence, by the triangle inequality,

$$\begin{aligned} f(P_2) &= d(P_1, P_2) + d(P_2, P_3) = d(P'_1, P_3) \\ &\leq d(P'_1, P') + d(P', P_3) \\ &\leq d(P_1, P) + d(P', P) + d(P, P_3) \\ &\leq d(P_1, P) + d(P, P_2) + d(P, P_3) = f(P), \end{aligned}$$

so P_2 is a global minimum of P_{FS} . □

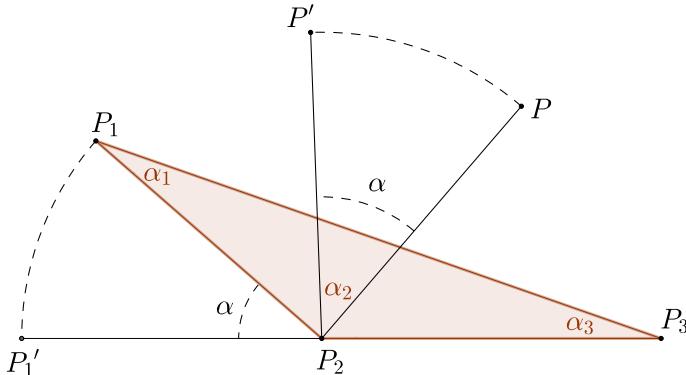


Fig. 4.3 Sketch of the proof of Proposition 4.7

The converse statement claiming that the attainment of the minimum at a vertex entails that $\max\{\alpha_1, \alpha_2, \alpha_3\} \geq 120$ degrees can also be proved geometrically [68, pp. 280–282].

From now on, we assume that $\max\{\alpha_1, \alpha_2, \alpha_3\} < 120$ degrees, i.e., that the minimum of f is not attained at one of the points P_1, P_2, P_3 . The first two solutions to P_{FS} under this assumption were obtained by Torricelli and his student Viviani, but were published in 1659, 12 years after the publication of Cavalieri's solution. Another interesting solution was provided by Steiner in 1842, which was later rediscovered by Gallai and, much later, by Hoffmann in 1929 (many results and methods have been rediscovered again and again before the creation of the first mathematical abstracting and reviewing services, Zentralblatt MATH., in 1930, and Mathematical Reviews, in 1940). Steiner's solution was based on successive 60 degree rotations in the plane, while Torricelli's one consisted in maximizing the area of those equilateral triangles whose sides contain exactly one of the points P_1, P_2 , and P_3 . A particular case of the latter problem was posed by Moss, in 1755, at the women's magazine *The Ladies Diary or the Woman's Almanack*, which suggests the existence of an important cohort of women with a good mathematical training during the Age of Enlightenment. The solution to the Torricelli–Moss problem, actually the dual problem to P_{FS} , was rediscovered by Vecten, in 1810, and by Fasbender, in 1846, to whom it was mistakenly attributed until 1991.

The second approach used to solve P_{FS} , called *analogical*, is inspired in the least action metaphysical principle asserting that nature always works in the most economic way, i.e., by minimizing certain magnitude: time in optics (as observed by Fermat), surface tension in soap bubbles, potential energy in gravitational fields, etc. The inconvenience of this analogical approach is that it is not easy to design an experiment involving some physical magnitude whose minimization is equivalent to the optimization problem to be solved. Moreover, nature does not compute global minima but local ones (in fact, critical points), as it happens with beads inserted in wires, whose equilibrium points are the local minima of the height (recall that the

potential energy of a particle of given mass m and variable height y is mgy , where g denotes the gravitational constant, so the gravitational field provides local minima of y on the curve represented by the wire). Moreover, these local minima might be attained in a parsimonious way, at least in the case of wildlife (evolution by natural selection is a very slow process).

The analogical approach has inspired many exotic numerical optimization methods imitating nature or even social behavior, but not always in a rigorous way. Such methods, called metaheuristic, are designed to generate, or select, a heuristic (partial search algorithm) that may provide a sufficiently good solution to a given optimization problem. In fact, “in recent years, the field of combinatorial optimization [where some decision variables take integer values] has witnessed a true tsunami of ‘novel’ metaheuristic methods, most of them based on a metaphor of some natural or man-made process. The behavior of virtually any species of insects, the flow of water, musicians playing together—it seems that no idea is too far-fetched to serve as inspiration to launch yet another metaheuristic” [82].

We illustrate the analogical approach by solving P_{FS} through two experiments that are based on the minimization of the surface tension of a soap bubble and the minimization of the potential energy of a set of mass points, respectively.

4.2.1.1 Minimizing the Surface Tension

The physicist Plateau witnessed, in 1840, a domestic accident caused by a servant who wrongly poured oil on a receptacle containing a mixture of alcohol and water. Intrigued by the spheric shape of the oil bubbles, Plateau started to perform various experiments with soapy water solutions, which lead him to the conclusion that the surface tension of the soap film is proportional to its surface area, so that the bubbles are spheric when they are at equilibrium (i.e., when the pressure of the air contained in the bubble equals the atmospheric pressure). Due to the lack of analytic tools, he could not justify his conjecture, which was proved in the twentieth century by Radó, in 1930, and by Douglas, in 1931, independently.

To analogically solve P_{FS} , it is sufficient to take two transparent plastic layers, a felt-tip pen, three sticks of equal length l , glue, and a soapy solution. Represent on both layers the three points P_1 , P_2 , and P_3 ; put both layers in parallel, at a distance l , so that each pair of points with the same label is vertically aligned, and link these pairs of points with the sticks and the glue. Submerging this frame in the soapy solution and getting it out, one obtains, at equilibrium, a minimal surface soap film linking the layers and the sticks. This surface is formed by rectangles of height l , so that the sum of areas is a local minimum for the area of perturbed soap films provided that the perturbations are sufficiently small (see Fig. 4.4). One of these equilibrium configurations corresponds to the triangular prism whose bases are the triangles of vertices P_1 , P_2 , and P_3 drawn on both layers, so in order to minimize the surface tension we may be obliged to repeat several times the experiment until we get the desired configuration, which provides the global minimum for P_{FS} on both layers.

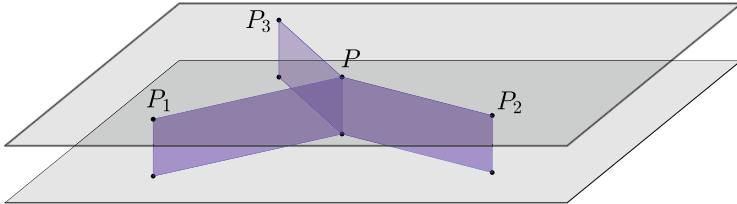


Fig. 4.4 Minimizing the surface tension

A similar construction allows to solve the Fermat–Steiner problem with more than three points.

4.2.1.2 Minimizing the Potential Energy

The famous Scottish Book was a notebook used in the 1930s and 1940s by mathematicians of the Lwów School of Mathematics for collecting interesting solved, unsolved, and even probably unsolvable problems (Lwów is the Polish name of the, at present, Ukrainian city of Lviv). The notebook was named after the Scottish Café where it was kept. Among the active participants at these meetings, there were the functional analysts and topologists Banach, Ulam, Mazur, and Steinhaus, who conceived an ingenious experiment to solve the weighted Fermat–Steiner problem. This problem seems to have been first considered in the calculus textbook published by Thomas Simpson in 1750. Tong and Chua (1995) have proposed a geometric solution inspired in Torricelli’s one for P_{FS} .

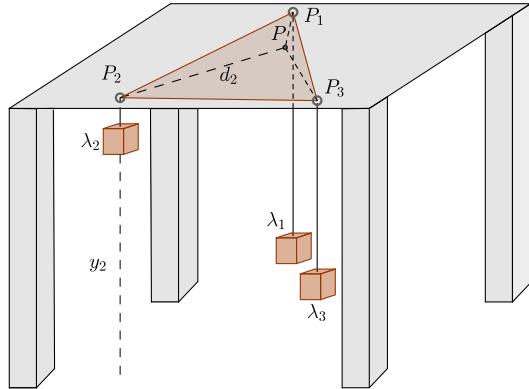
Steinhaus’ experiment requires a table, whose height will be taken as unit length, a drill and three pieces of string of a unit of length tied at one of their extremes. We assume that the three sides of the triangle of vertices P_1 , P_2 , and P_3 have length less than 1 (a rescaling may be necessary to get this assumption). Draw the three points P_1 , P_2 , and P_3 on the table, drill a hole at each of these points, and introduce one of the strings through each of the holes. Tie a small solid of mass λ_i at the string hanging from P_i , $i = 1, 2, 3$ (see Fig. 4.5).

Assume that the knot is at P , placed at a distance $d_i = d(P, P_i)$ from P_i , and denote by y_i the height to the floor of the i th solid. Then, since $y_i + 1 - d_i = 1$ (i.e., $y_i = d_i$), the potential energy of the three-mass system is the product of the gravitational constant g times

$$\sum_{i=1}^3 \lambda_i y_i = \sum_{i=1}^3 \lambda_i d(P, P_i).$$

Taking $\lambda_1 = \lambda_2 = \lambda_3$, the potential energy is proportional to $f(P) := d(P, P_1) + d(P, P_2) + d(P, P_3)$, subject to some constraints that do not need to be explicitly stated while P lies in the interior of the triangle determined by P_1 , P_2 , and P_3 , as

Fig. 4.5 Minimizing the gravitational potential energy



they are not active at the optimal solution. To prove the existence and uniqueness of optimal solution, one needs an analytical argument based on coercivity and convexity. Nothing changes when one considers more than three points, whether weighted or not.

4.2.1.3 The Analytical Solution

Consider the plane equipped with rectangular coordinate axis. We can identify the four points P_1, P_2, P_3 , and P with $x_1, x_2, x_3, x \in \mathbb{R}^2$, and P_{FS} is then formulated as:

$$P_{FS} : \text{Min}_{x \in \mathbb{R}^2} f(x) = \sum_{i=1}^3 \|x - x_i\|.$$

We assume that x_1, x_2, x_3 are not aligned and that the size of the angles of the triangle with vertices x_1, x_2, x_3 is all less than 120 degrees.

We start by exploring the properties of the objective function $f = \sum_{i=1}^3 f_i$, where $f_i(x) = \|x - x_i\|$. Obviously, f_i is convex as it is the result of composing an affine function, $x \mapsto x - x_i$, and a convex one (the norm), $i = 1, 2, 3$, so f is convex too. Thanks to the convexity of P_{FS} , the local minima obtained via the analogical approach are actually global minima. We now prove that f is strictly convex and coercive.

We first show the strict convexity of f in an intuitive way. Consider the norm function, $h(x) := \|x\|$, whose graph, $\text{gph } h$, is the ice-cream cone of Fig. 4.6. Given $x, d \in \mathbb{R}^2$, with $d \neq 0_2$, the graph of the one variable real-valued function $h_{x,d}$ given by $h_{x,d}(t) = \|x + td\|$ (the section of h determined by x and d) is a branch of hyperbola if $x \notin \text{span}\{d\}$, and it is the union of two half-lines if $x \in \text{span}\{d\}$. So, $h_{x,d}$ is strictly convex if and only if $x \notin \text{span}\{d\}$. Consequently, the section of f_i determined by the line containing x in the direction of d is strictly convex if and only if $x - x_i \notin \text{span}\{d\}$. Since the sum of two convex functions is strictly convex when one of them is strictly convex, all sections of $f_1 + f_2$ are strictly convex except in

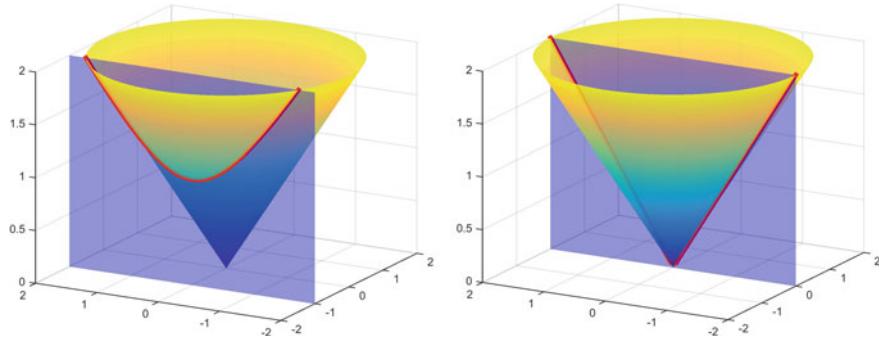


Fig. 4.6 Intersection of $\text{gph } \|\cdot\|$ with vertical planes

the case that the line containing x in the direction of d contains x_1 and x_2 . Finally, $f = \sum_{i=1}^3 f_i$ is strictly convex due to the assumptions that x_1 , x_2 , and x_3 are not aligned, so at least one of the three (convex) sections is strictly convex.

We now prove the strict convexity of f analytically, by contradiction, exploiting the fact that, given two vectors $u, v \in \mathbb{R}^2 \setminus \{0_2\}$, if $\|u + v\| = \|u\| + \|v\|$, then $u^T v = \|u\| \|v\|$, and so, there exists $\mu > 0$ such that $u = \mu v$. Suppose that f is not strictly convex. Let $y, z \in \mathbb{R}^2$, $y \neq z$ and $\lambda \in]0, 1[$, be such that

$$f((1 - \lambda)y + \lambda z) = (1 - \lambda)f(y) + \lambda f(z),$$

i.e.,

$$\sum_{i=1}^3 f_i((1 - \lambda)y + \lambda z) = (1 - \lambda) \sum_{i=1}^3 f_i(y) + \lambda \sum_{i=1}^3 f_i(z).$$

Then, due to the convexity of f_i , $i = 1, 2, 3$, one necessarily has

$$f_i((1 - \lambda)y + \lambda z) = (1 - \lambda)f_i(y) + \lambda f_i(z), \quad i = 1, 2, 3,$$

or, equivalently,

$$\|(1 - \lambda)(y - x_i) + \lambda(z - x_i)\| = \|(1 - \lambda)(y - x_i)\| + \|\lambda(z - x_i)\|, \quad i = 1, 2, 3. \quad (4.4)$$

Let $i \in \{1, 2, 3\}$. If $y \neq x_i \neq z$, one has $(1 - \lambda)(y - x_i), \lambda(z - x_i) \in \mathbb{R}^2 \setminus \{0_2\}$ and (4.4) implies the existence of $\mu_i > 0$ such that $(1 - \lambda)(y - x_i) = \mu_i \lambda(z - x_i)$. Defining $\gamma_i := \frac{\mu_i \lambda}{1 - \lambda} > 0$, we have $y - x_i = \gamma_i(z - x_i)$, with $\gamma_i \neq 1$ as $y \neq z$. Hence,

$$x_i = \left(\frac{1}{1 - \gamma_i} \right) y - \left(\frac{\gamma_i}{1 - \gamma_i} \right) z \in L,$$

where L is the line containing y and z . Alternatively, if $y \neq x_i \neq z$ does not hold, we have $x_i \in \{y, z\} \subset L$. We conclude that $\{x_1, x_2, x_3\} \subset L$; i.e., the points x_1, x_2 , and x_3 are aligned (contradiction).

The functions f_i , $i = 1, 2, 3$, are coercive as $f_i(x) = \|x - x_i\| \geq \|x\| - \|x_i\|$ implies that

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

Thus, their sum f is also coercive. In summary, P_{FS} has a unique optimal solution. From this fact, and the observation that $\nabla f_i(x) = \frac{x - x_i}{\|x - x_i\|}$ for all $x \neq x_i$, $i = 1, 2, 3$, we deduce the following result.

Proposition 4.8 (Analytic solution to the Fermat–Steiner problem) *There exists a unique global minimizer of f . Moreover, such a global minimizer is the element $x \in \mathbb{R}^2 \setminus \{0_2\}$ such that*

$$\nabla f(x) = \sum_{i=1}^3 \frac{x - x_i}{\|x - x_i\|} = 0_2. \quad (4.5)$$

Proposition 4.9 (Geometric solution to the Fermat–Steiner problem) *The optimal solution of P_{FS} is the isogonic center of the triangle with vertices x_1, x_2 , and x_3 , that is, the point \bar{x} from which the three sides are seen under the same angle of 120 degrees.*

Proof Let α_{12}, α_{13} , and α_{23} be the angles under which the sides $[x_1, x_2]$, $[x_1, x_3]$, and $[x_2, x_3]$ are seen by an observer placed at x satisfying (4.5). Since f is differentiable at x , the Fermat necessary optimality condition yields $\nabla f(x) = 0_2$, i.e.,

$$u + v + w = 0_2, \quad (4.6)$$

where $u := \frac{x - x_1}{\|x - x_1\|}$, $v := \frac{x - x_2}{\|x - x_2\|}$ and $w := \frac{x - x_3}{\|x - x_3\|}$. Since $\|u\| = \|v\| = \|w\| = 1$, $u^T v = \cos \alpha_{12}$, $u^T w = \cos \alpha_{13}$, and $v^T w = \cos \alpha_{23}$.

By multiplying both sides of (4.6) by u , v , and w , one gets a linear system of equations whose unknowns are the cosines of the angles α_{12}, α_{13} , and α_{23} ,

$$\begin{cases} 1 + \cos \alpha_{12} + \cos \alpha_{13} = 0 \\ \cos \alpha_{12} + 1 + \cos \alpha_{23} = 0 \\ \cos \alpha_{13} + \cos \alpha_{23} + 1 = 0 \end{cases},$$

whose unique solution is

$$\cos \alpha_{12} = \cos \alpha_{13} = \cos \alpha_{23} = -\frac{1}{2},$$

that is,

$$\alpha_{12} = \alpha_{13} = \alpha_{23} = 120 \text{ degrees}.$$

Hence, the unique solution to $\nabla f(x) = 0_2$ is the isogonic center \bar{x} . \square

Example 4.10 ([68, pp. 284–285]). In the 1960s, Bell Telephone installed telephone networks connecting the different headquarters of the customer companies with a cost, regulated by law, which was proportional to the total length of the installed network. One of these customers, Delta Airlines, wanted to connect its hubs placed at the airports of Atlanta, Chicago, and New York, which approximately formed an equilateral triangle. Bell Telephone proposed to install cables connecting one of the bases with the other two bases, but Delta Airlines proposed instead to create a virtual base at the isogonic center of the triangle formed by the three hubs, linking this point with the three bases by cables. The Federal Authority agreed with this solution, and Delta Airlines earned a significant amount of money that we can estimate. Take the average distance between hubs as length unit. Since the height of the equilateral triangle is $\frac{\sqrt{3}}{2}$, the distance from the isogonic center (here coinciding with the orthocenter) to any vertex is $\frac{2}{3} \times \frac{\sqrt{3}}{2} = \frac{\sqrt{3}}{3}$. Therefore, the length of the networks proposed by Bell Telephone and by Delta Airlines was of 2 and $\sqrt{3}$ units, with an estimated earning of $\left(\frac{2-\sqrt{3}}{2}\right) \times 100 = 13.4\%$.

4.2.2 The Fixed-Point Method of Weiszfeld

A specific numerical method for P_{FS} was proposed by Weiszfeld in 1937, when he was only 16. Indeed, since

$$\sum_{i=1}^3 \frac{\bar{x} - x_i}{\|\bar{x} - x_i\|} = 0_2$$

is equivalent to

$$\left(\sum_{i=1}^3 \frac{1}{\|\bar{x} - x_i\|} \right) \bar{x} = \left(\sum_{i=1}^3 \frac{x_i}{\|\bar{x} - x_i\|} \right),$$

that is,

$$\bar{x} = \frac{\left(\sum_{i=1}^3 \frac{x_i}{\|\bar{x} - x_i\|} \right)}{\left(\sum_{i=1}^3 \frac{1}{\|\bar{x} - x_i\|} \right)},$$

the optimality condition (4.5) is equivalent to assert that \bar{x} is a *fixed point* for the function

$$h(x) := \frac{\left(\sum_{i=1}^3 \frac{x_i}{\|x-x_i\|} \right)}{\left(\sum_{i=1}^3 \frac{1}{\|x-x_i\|} \right)}.$$

Inspired in the argument of the Banach fixed-point theorem, proved in 1922, Weiszfeld considered sequences $\{x^k\} \subset \mathbb{R}^3$ whose initial element (seed) $x^0 \in \mathbb{R}^2 \setminus \{x_1, x_2, x_3\}$ is arbitrary, and $x^{k+1} = h(x^k)$, $k = 1, 2, \dots$. Of, course, $\{x^k\}$ could be finite as h fails to be differentiable at three points and, in the best case that it is infinite, it does not necessarily converges as h is not contractive. Harold W. Kuhn rediscovered Weiszfeld's method in 1973 and showed that $\{x^k\}$ always converges whenever it is infinite. Recent research has shown that the set of bad seeds, formed by the initial elements for which $\{x^k\}$ attains an element of $\{x_1, x_2, x_3\}$ after a finite number of steps, is countable [9, 48]. Since the set of bad seeds is very small (more precisely, it is a Lebesgue zero-measure set), one concludes that the sequences generated by Weiszfeld's method converge to the isogonic point with probability one when the seeds are selected at random in the triangle. A similar behavior has been observed, in practice, for all the general optimization methods described in Chap. 5, including those which assume the existence of gradients or Hessian matrices of f .

4.2.3 An Application to Statistical Inference

One of the main aims of statistical inference consists in estimating the parameters of the density function of a scalar random variable x , which is assumed to belong to a given family, from a sample $(x_1, \dots, x_n) \in \mathbb{R}^n$ (the result of $n \geq 2$ realizations of x). The most common method to solve this statistical inference problem is the maximum likelihood one, which consists in computing the unique global maximum $\bar{\theta}$ of the *maximum likelihood function* of the sample, that is,

$$f(x_1, \dots, x_n \mid \theta) := \prod_{i=1}^n f(x_i \mid \theta),$$

where $f(x \mid \theta)$ represents the density function of x and θ is the parameter to be estimated.

Assume, for instance, that x has a *normal distribution* of parameter $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{++}$, i.e., that the density function of x is

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right),$$

so we have that the maximum likelihood function is, in this case,

$$f(x_1, \dots, x_n | \mu, \sigma^2) := \prod_{i=1}^n f(x_i | \mu, \sigma^2) = \left(\frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} \exp\left(-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right).$$

The optimization problem to be solved is then

$$\begin{aligned} P_1 : \text{Max } & f(x_1, \dots, x_n | \mu, \sigma^2) \\ \text{s.t. } & (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{++}, \end{aligned}$$

which can be reformulated, recalling that $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$ is the *sample mean* while $s^2 := \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ is the *sample variance*. We first observe that $(\frac{1}{2\pi})^{\frac{n}{2}}$ is constant, while

$$\begin{aligned} \sum_{i=1}^n (x_i - \mu)^2 &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu)^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + 2(\bar{x} - \mu) \sum_{i=1}^n (x_i - \bar{x}) + n(\bar{x} - \mu)^2 \\ &= n[s^2 + (\bar{x} - \mu)^2]. \end{aligned}$$

Replacing in P_1 the (positive) objective function by its natural logarithm, changing the task “Max” by the task “Min” (i.e., changing the sign of the objective function), and making the change of variable $\delta := \sigma^2 > 0$, one gets the equivalent problem:

$$\begin{aligned} P_2 : \text{Min } & g(\mu, \delta) := \frac{1}{2} \ln \delta + \frac{1}{2\delta} [s^2 + (\bar{x} - \mu)^2] \\ \text{s.t. } & (\mu, \delta) \in \mathbb{R} \times \mathbb{R}_{++}, \end{aligned}$$

whose constraint set $\mathbb{R} \times \mathbb{R}_{++}$ is open and convex, and $g \in \mathcal{C}^2(\mathbb{R} \times \mathbb{R}_{++})$; see Fig. 4.7.

Since $\text{dom } g = \mathbb{R} \times \mathbb{R}_{++}$ is not closed, we cannot derive the existence of a global minimum via coercivity. Fortunately, we can prove this existence by a suitable convexity argument. We have

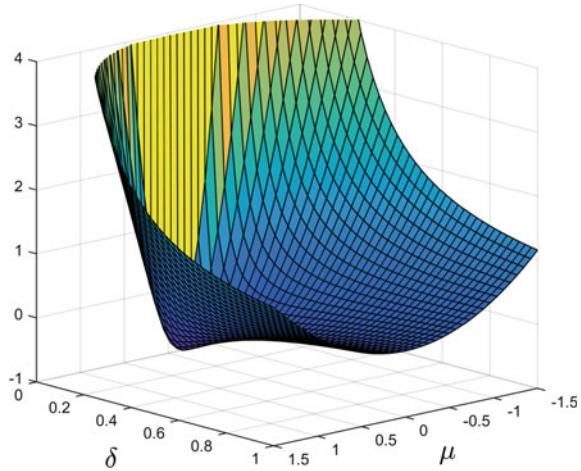
$$\frac{\partial g}{\partial \mu} = -\frac{\bar{x} - \mu}{\delta},$$

$$\frac{\partial g}{\partial \delta} = \frac{1}{2\delta} - \frac{1}{2\delta^2} (s^2 + (\bar{x} - \mu)^2),$$

and

$$\nabla^2 g(\mu, \delta) = \begin{bmatrix} \frac{1}{\delta} & \frac{\bar{x} - \mu}{\delta^2} \\ \frac{\bar{x} - \mu}{\delta^2} & \frac{1}{\delta^3} (s^2 + (\bar{x} - \mu)^2) - \frac{1}{2\delta^2} \end{bmatrix},$$

Fig. 4.7 $\text{gph } g$ for $\bar{x} = 0$
and $s = \frac{1}{4}$



with

$$\det \nabla^2 g(\mu, \delta) = \frac{1}{\delta^4} (s^2 + (\bar{x} - \mu)^2) - \frac{1}{2\delta^3} - \frac{(\bar{x} - \mu)^2}{\delta^4} = \frac{s^2}{\delta^4} - \frac{1}{2\delta^3}.$$

Thus, $\nabla^2 g$ is positive definite on the open convex set $C := \mathbb{R} \times]0, 2s^2[$. Since g is convex and differentiable on C , the minima of g on C are its critical points. Obviously, $\frac{\partial g}{\partial \mu} = 0$ if and only if $\mu = \bar{x}$. Then,

$$\frac{\partial g}{\partial \delta} = \frac{1}{2\delta} - \frac{1}{2\delta^2} (s^2 + (\bar{x} - \mu)^2) = \frac{1}{2\delta} - \frac{s^2}{2\delta^2} = 0$$

if and only if $\delta = s^2$, so we have that

$$(\hat{\mu}, \hat{\delta}) := (\bar{x}, s^2)$$

is the unique global minimum of g on C , by Proposition 4.1(v). In particular,

$$g(\hat{\mu}, \hat{\delta}) \leq g(\mu, \delta), \quad \forall \mu \in \mathbb{R}, \forall \delta \leq \hat{\delta}.$$

It remains to prove that

$$g(\hat{\mu}, \hat{\delta}) \leq g(\mu, \delta), \quad \forall \mu \in \mathbb{R}, \forall \delta > \hat{\delta},$$

or, equivalently, since $\inf \{g(\mu, \delta) : \mu \in \mathbb{R}, \delta > \hat{\delta}\} = g(\bar{x}, \delta)$,

$$g(\bar{x}, \hat{\delta}) \leq g(\bar{x}, \delta), \quad \forall \delta > \hat{\delta},$$

as

$$g(\mu, \delta) = \frac{1}{2} \ln \delta + \frac{1}{2\delta} (s^2 + (\bar{x} - \mu)^2) \geq g(\bar{x}, \delta), \quad \forall \delta > 0.$$

Consider the function $h(\delta) := g(\bar{x}, \delta) = \frac{1}{2} \ln \delta + \frac{s^2}{2\delta}$. Since

$$h'(\delta) = \frac{1}{2\delta} - \frac{s^2}{2\delta^2} \geq 0, \quad \forall \delta \in [s^2, +\infty[= [\hat{\delta}, +\infty[,$$

one has

$$h(\hat{\delta}) \leq h(\delta), \quad \forall \delta \geq \hat{\delta},$$

so $(\hat{\mu}, \hat{\delta})$ is a global minimum of g on $\mathbb{R} \times \mathbb{R}_{++}$. The uniqueness follows from the fact that $(\hat{\mu}, \hat{\delta})$ is the unique critical point of g on $\mathbb{R} \times \mathbb{R}_{++}$. We thus get the aimed conclusion:

Proposition 4.11 (Maximum likelihood estimators) *The maximum likelihood estimators of the parameters μ and σ^2 of a normally distributed random variable are the sample mean and the variance, respectively.*

4.3 Linearly Constrained Convex Optimization

We consider in this section convex optimization problems of the form

$$\begin{aligned} P : \text{Min } f(x) \\ \text{s.t. } g_i(x) \leq 0, \quad i \in I := \{1, \dots, m\}, \end{aligned}$$

where $g_i(x) = a_i^T x - b_i$, $i \in I$, and f is convex and differentiable on an open convex set C that contains the feasible set, denoted by F .

4.3.1 Optimality Conditions

Definition 4.12 We say that $d \in \mathbb{R}^n$ is a *feasible direction* at $\bar{x} \in F$ if there exists $\varepsilon > 0$ such that $\bar{x} + td \in F$ for all $t \in [0, \varepsilon]$ (i.e., if we can move from \bar{x} in the direction d without leaving F). The set of feasible directions at \bar{x} form a convex cone called *feasible direction cone*, which we denote by $D(\bar{x})$, i.e.,

$$D(\bar{x}) := \{d \in \mathbb{R}^n : \exists \varepsilon > 0 \text{ such that } \bar{x} + td \in F, \forall t \in [0, \varepsilon]\},$$

We denote by $I(\bar{x})$ the *set of active indices* at $\bar{x} \in F$, that is,

$$I(\bar{x}) := \{i \in I : g_i(\bar{x}) = 0\}.$$

We define the *active cone* at $\bar{x} \in F$ as the convex cone generated by the gradients of the active constraints at that point, i.e.,

$$A(\bar{x}) := \text{cone}\{\nabla g_i(\bar{x}) : i \in I(\bar{x})\}. \quad (4.7)$$

In other words,

$$I(\bar{x}) = \{i \in I : a_i^T \bar{x} = b_i\} \quad \text{and} \quad A(\bar{x}) = \text{cone}\{a_i : i \in I(\bar{x})\}.$$

Since $I(\bar{x})$ is finite, $A(\bar{x})$ is a finitely generated convex cone, and so, it is closed (see Proposition 2.13).

Definition 4.13 The *(negative) polar cone* of a nonempty set $Y \subset \mathbb{R}^n$ is defined as

$$Y^\circ := \{z \in \mathbb{R}^n : y^T z \leq 0, \forall y \in Y\}.$$

By convention, the polar cone of the empty set \emptyset is the whole space \mathbb{R}^n .

Obviously, the polar cone is a closed convex cone (even if Y is not convex). It is easy to see that

$$A(\bar{x})^\circ = \{y \in \mathbb{R}^n : a^T y \leq 0, \forall a \in A(\bar{x})\} = \{y \in \mathbb{R}^n : a_i^T y \leq 0, \forall i \in I(\bar{x})\}.$$

Since $I(\bar{x})$ is finite, $A(\bar{x})^\circ$ is a polyhedral cone.

Proposition 4.14 (Computing the feasible direction cone) *The feasible direction cone at $\bar{x} \in F$ satisfies the equation*

$$D(\bar{x}) = A(\bar{x})^\circ. \quad (4.8)$$

Proof We shall prove both inclusions.

Let $d \in D(\bar{x})$ and $\varepsilon > 0$ be such that $\bar{x} + t d \in F$ for all $t \in [0, \varepsilon]$. For any $i \in I(\bar{x})$, one has $a_i^T \bar{x} = b_i$, so we have that

$$b_i \geq a_i^T (\bar{x} + \varepsilon d) = a_i^T \bar{x} + \varepsilon a_i^T d = b_i + \varepsilon a_i^T d.$$

Hence, $a_i^T d \leq 0$ for all $i \in I(\bar{x})$. Thus, $d \in A(\bar{x})^\circ$.

To prove the reverse inclusion, take an arbitrary $d \in A(\bar{x})^\circ$. If $i \notin I(\bar{x})$, then $a_i^T \bar{x} < b_i$, so there exists a constant $\eta > 0$ such that

$$a_i^T \bar{x} + \eta < b_i, \quad \forall i \notin I(\bar{x}).$$

Moreover, there exists a constant $\varepsilon > 0$ such that $\varepsilon |a_i^T d| < \eta$ for all $i \in I$. Take now any $t \in [0, \varepsilon]$ and any $i \in I$.

If $i \in I(\bar{x})$, as $d \in A(\bar{x})^\circ$, then

$$a_i^T(\bar{x} + td) = b_i + ta_i^T d \leq b_i.$$

If $i \notin I(\bar{x})$, then

$$a_i^T(\bar{x} + td) = a_i^T\bar{x} + ta_i^T d \leq a_i^T\bar{x} + t|a_i^T d| \leq a_i^T\bar{x} + \varepsilon|a_i^T d| \leq a_i^T\bar{x} + \eta < b_i,$$

so we have that $a_i^T(\bar{x} + td) \leq b_i$, for all $i \in I$. Hence, $\bar{x} + td \in F$ for all $t \in [0, \varepsilon]$, that is, $d \in D(\bar{x})$. \square

Lemma 4.15 (Generalized Farkas lemma) *For any $Y \subset \mathbb{R}^n$, it holds*

$$Y^{\circ\circ} = \text{cl cone } Y.$$

Proof Let $y \in Y^{\circ\circ}$, and let us assume, by contradiction, that $y \notin \text{cl cone } Y$. Since $\text{cl cone } Y$ is a closed convex cone, by the separation theorem for convex cones (Corollary 2.15), there exists $a \in \mathbb{R}^n$ such that $a^T y > 0$ and

$$a^T x \leq 0, \quad \forall x \in \text{cl cone } Y. \tag{4.9}$$

From (4.9), one gets that $a \in Y^\circ$. Since $y \in Y^{\circ\circ}$, we should have $a^T y \leq 0$, getting in this way the aimed contradiction.

To prove the reverse inclusion, we now show that $Y \subset Y^{\circ\circ}$ (from this inclusion, one concludes that $\text{cl cone } Y \subset \text{cl cone } Y^{\circ\circ} = Y^{\circ\circ}$, as $Y^{\circ\circ}$ is a closed convex cone). Take any $y \in Y$. By the definition of Y° , one has $x^T y \leq 0$ for all $x \in Y^\circ$, and so, $y \in Y^{\circ\circ}$. \square

We next characterize those homogeneous linear inequalities which are satisfied by all the solutions to a given homogeneous linear inequality system. In two dimensions, this result characterizes those half-planes which contain a given angle with apex at the origin and whose boundaries go through the origin.

Corollary 4.16 (Generalized Farkas lemma for linear inequalities) *Consider the system $\{a_i^T x \leq 0, i \in I\}$ of linear inequalities in \mathbb{R}^n , with I being an arbitrary (possibly infinite) index set. Then,*

$$\forall x \in \mathbb{R}^n : [a_i^T x \leq 0, \forall i \in I] \Rightarrow [a^T x \leq 0]$$

if and only if

$$a \in \text{cl cone}\{a_i, i \in I\}.$$

Proof This is a direct application of Lemma 4.15 to the set $Y := \{a_i, i \in I\}$. Indeed, $a \in Y^{\circ\circ}$, by definition, if $a^T x \leq 0$ for all $x \in \mathbb{R}^n$ such that $a_i^T x \leq 0$ for all $i \in I$. By Lemma 4.15, we have that $Y^{\circ\circ} = \text{cl cone}\{a_i, i \in I\}$, which proves the claim. \square

Trying to characterize in a rigorous way the equilibrium points of dynamic systems, the Hungarian physicist G. Farkas proved in 1901, after several failed attempts,

the particular case of the above result where the index set I is finite and the closure operator can be removed (recall that any finitely generated convex cone is closed). This *classical Farkas lemma* will be used in this chapter to characterize optimal solutions to convex problems and to linearize conic systems. Among its many applications, let us mention machine learning [63], probability, economics, and finance [32, 37]. The generalized Farkas lemma (whose infinite dimensional version was proved by Chu [22]) is just one of the many available extensions of the classical Farkas lemma [31]. We use this version in Chapter 6 to obtain necessary optimality conditions in nonconvex optimization.

The following proposition lists some basic properties of polar cones to be used hereinafter.

Proposition 4.17 (Handling polar sets) *Given $Y, Z \subset \mathbb{R}^n$, the following properties hold:*

- (i) *If $Y \subset Z$, then $Z^\circ \subset Y^\circ$.*
- (ii) *$Y^\circ = (\text{cone } Y)^\circ = (\text{cl cone } Y)^\circ$.*
- (iii) *$Y^{\circ\circ} = Y$ if and only if Y is a closed convex cone.*

Proof Statements (i) and (ii) come straightforwardly from the definition of polar cone, while (iii) is immediate from the Farkas Lemma 4.15. \square

Corollary 4.18 *The feasible direction cone at $\bar{x} \in F$ satisfies the equation*

$$D(\bar{x})^\circ = A(\bar{x}). \quad (4.10)$$

Proof Taking the polar at both members of (4.8), and applying Lemma 4.15, one gets

$$D(\bar{x})^\circ = A(\bar{x})^{\circ\circ} = \text{cl } A(\bar{x}) = A(\bar{x}),$$

as $A(\bar{x})$ is a closed convex cone. \square

Example 4.19 Consider the polyhedral set

$$F := \{x \in \mathbb{R}^2 : -2x_1 + x_2 \leq 0, -x_1 - x_2 \leq 0, -x_2 - 4 \leq 0\}$$

and the point $\bar{x} = 0_2 \in F$. We have $I(\bar{x}) = \{1, 2\}$ and

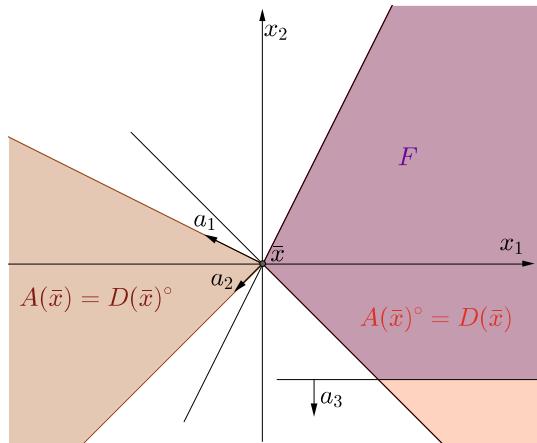
$$A(\bar{x}) = \text{cone} \left\{ \begin{pmatrix} -2 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix} \right\}.$$

The feasible direction cone at \bar{x} can be expressed as

$$D(\bar{x}) = \{x \in \mathbb{R}^2 : -2x_1 + x_2 \leq 0, -x_1 - x_2 \leq 0\}.$$

We can easily check that $D(\bar{x})^\circ = A(\bar{x})$ and $A(\bar{x})^\circ = D(\bar{x})$ (see Fig. 4.8).

Fig. 4.8 The feasible direction cone and the active cone are polar of each other



We now prove, from the previously obtained relationships between the feasible direction cone and the active cone, the simplest version of the Karush–Kuhn–Tucker (KKT in brief) theorem, which provides a first-order characterization for linearly constrained convex optimization problems in terms of the so-called *nonnegativity condition* (NC), the *stationarity condition* (SC), and the *complementarity condition* (CC), altogether called the *KKT conditions*. The general version of the KKT theorem, which provides a necessary optimality condition for nonlinear optimization problems with inequality constraints, was first proved by Karush in his unpublished master's thesis on the extension of the method of Lagrange multipliers for equality constrained problems, in 1939, and rediscovered by Kuhn and Tucker in 1951.

Theorem 4.20 (KKT theorem with linear constraints) *Let $\bar{x} \in F$. Then, the following statements are equivalent:*

- (i) $\bar{x} \in F^*$.
- (ii) $-\nabla f(\bar{x}) \in A(\bar{x})$.
- (iii) There exists $\bar{\lambda} \in \mathbb{R}^m$ such that:

$$\begin{aligned} (NC) \quad & \bar{\lambda} \in \mathbb{R}_+^m; \\ (SC) \quad & \nabla f(\bar{x}) + \sum_{i \in I} \bar{\lambda}_i a_i = 0_n; \text{ and} \\ (CC) \quad & \bar{\lambda}_i (b_i - a_i^T \bar{x}) = 0, \quad \forall i \in I. \end{aligned}$$

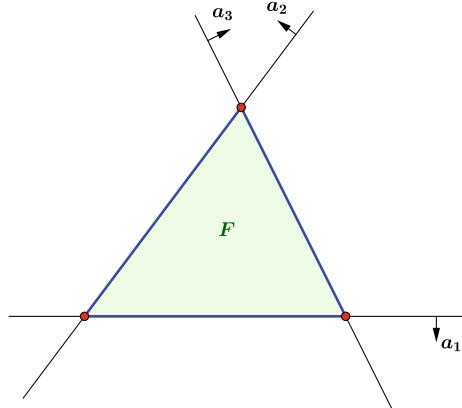
Proof We shall prove that (i) \Leftrightarrow (ii) and (ii) \Leftrightarrow (iii).

(i) \Rightarrow (ii) Assume that $\bar{x} \in F^*$. Let $d \in D(\bar{x})$ and $\varepsilon > 0$ be such that $\bar{x} + td \in F$ for all $t \in [0, \varepsilon]$. Since $f(\bar{x} + td) \geq f(\bar{x})$ for all $t \in [0, \varepsilon]$,

$$\nabla f(\bar{x})^T d = f'(\bar{x}; d) = \lim_{t \searrow 0} \frac{f(\bar{x} + td) - f(\bar{x})}{t} \geq 0.$$

Then, recalling (4.10), one gets

$$-\nabla f(\bar{x}) \in D(\bar{x})^\circ = A(\bar{x}).$$

Fig. 4.9 Partition of F 

(ii) \Rightarrow (i) Assume now that $\bar{x} \in F$ satisfies $-\nabla f(\bar{x}) \in A(\bar{x})$, that is, $-\nabla f(\bar{x}) \in D(\bar{x})^\circ$. Let $x \in F$. Since $x - \bar{x} \in D(\bar{x})$, we have $\nabla f(\bar{x})^T(x - \bar{x}) \geq 0$. Then, due to the convexity of f (see Proposition 2.40),

$$f(x) \geq f(\bar{x}) + \nabla f(\bar{x})^T(x - \bar{x}) \geq f(\bar{x}),$$

so we have that $\bar{x} \in F^*$.

(ii) \Leftrightarrow (iii) Let $-\nabla f(\bar{x}) \in A(\bar{x})$, i.e.,

$$-\nabla f(\bar{x}) \in \text{cone}\{a_i, i \in I(\bar{x})\}.$$

Then, there exist scalars $\bar{\lambda}_i \geq 0$, $i \in I(\bar{x})$, such that $-\nabla f(\bar{x}) = \sum_{i \in I(\bar{x})} \bar{\lambda}_i a_i$. Defining $\bar{\lambda}_i = 0$ for all $i \in I \setminus I(\bar{x})$, the KKT vector $\bar{\lambda} := (\bar{\lambda}_1, \dots, \bar{\lambda}_m)^T \in \mathbb{R}^m$ satisfies the three conditions: (NC), (SC), and (CC).

The converse statement is trivial. \square

The KKT conditions allow to solve occasionally, with pen and paper, small convex optimization problems by enumerating the feasible solutions which could be optimal at any of the 2^m (possibly empty) subsets of F which result in discussing the possible values of the index set $I(x) \subset \{1, \dots, m\}$. For instance, if F is a triangle in \mathbb{R}^2 described by a linear system $\{a_i^T x \leq b_i, i \in I\}$, then $I(x) = \emptyset$ for $x \in \text{int } F$ (the green set in Fig. 4.9), $|I(x)| = 2$ for the three vertices (the red points), and $|I(x)| = 1$ for the points on the three sides which are not vertices (the blue segments). Observe that, since there is no $x \in F$ such that $|I(x)| = 3$, the $7 = 2^3 - 1$ mentioned subsets of F provide a partition of the feasible set.

Once the set $I(x)$ has been fixed, the optimal solutions in the corresponding subset of F are the x -part of the pairs $(x, \lambda) \in \mathbb{R}^n \times \mathbb{R}^{|I(x)|}$ such that (NC) and (SC) hold, i.e.,

$$\begin{aligned} -\nabla f(x) &= \sum_{i \in I(x)} \lambda_i a_i, \\ a_i^T x &= b_i, \quad \forall i \in I(x), \\ \lambda_i &\geq 0, \quad \forall i \in I(x), \\ a_i^T x &< b_i, \quad \forall i \in I \setminus I(x). \end{aligned} \tag{4.11}$$

The difficulty of solving the subsystem of $n + |I(x)|$ equations and unknowns from the *KKT system* (4.11) derives from the nonlinear nature of $\nabla f(x)$. A solution (x, λ) is accepted when the m inequalities in (4.11) are satisfied.

In the presence of equations, the mentioned discussion on $I(x)$ only affects the inequality constraints. For instance, if the constraints of P are $a_1^T x = b_1$ and $a_2^T x \leq b_2$, the possible values of $I(x)$ are $\{1\}$ and $\{1, 2\}$, with associated KKT systems

$$\{-\nabla f(x) = \lambda_1 a_1, a_1^T x = b_1, a_2^T x < b_2\}$$

and

$$\{-\nabla f(x) = \lambda_1 a_1 + \lambda_2 a_2, \lambda_2 \geq 0, a_1^T x = b_1, a_2^T x = b_2\},$$

respectively.

Example 4.21 The design problem in Example 1.6 was first formulated as

$$\begin{aligned} P_1 : \text{Min } f_1(x) &= x_1 x_2 + x_1 x_3 + x_2 x_3 \\ \text{s.t. } x_1 x_2 x_3 &= 1, \\ x &\in \mathbb{R}_{++}^3. \end{aligned}$$

The change of variables $y_i := \ln x_i, i = 1, 2, 3$, provides the equivalent linearly constrained convex problem

$$\begin{aligned} P_2 : \text{Min } f_2(y) &= e^{y_1+y_2} + e^{y_1+y_3} + e^{y_2+y_3} \\ \text{s.t. } y_1 + y_2 + y_3 &= 0. \end{aligned}$$

The unique point satisfying the KKT conditions is 0_3 , with KKT vectors $(\bar{\lambda}_1, \bar{\lambda}_2)^T$ such that $\bar{\lambda}_1 - \bar{\lambda}_2 = 2$. Then, $F_2^* = \{0_3\}$ and thus $F_1^* = \{(1, 1, 1)^T\}$. This shows that we may have a unique optimal solution to P with infinitely many corresponding KKT vectors.

In the particular case of linear optimization, where $f(x) = c^T x$ for some vector $c \in \mathbb{R}^n$, the global minima are characterized by the condition $-c \in A(\bar{x})$.

It is worth observing that we have proved (i) \Rightarrow (ii) \Leftrightarrow (iii) in Theorem 4.20 under the unique assumption that \bar{x} is a local minimum of the problem. So, if f is not convex, (ii) and the equivalent statement (iii) are necessary conditions for \bar{x} to be a local minimum. The application of this useful necessary condition allows to filter the candidates for local minima and for computing the global minima whenever F is compact or f is coercive on F (by comparing candidates). Another useful trick consists in tackling, instead of the given problem, the result of eliminating some

constraints, called *relaxed problem*, whose feasible set is generally greater than the initial one (this is systematically made when the constraint set C is discrete, e.g., when the decision variables are integer, i.e., $C = \mathbb{Z}^n$). The next example illustrates the application of both tools to solve nonconvex optimization problems.

Example 4.22 The objective function f of

$$\begin{aligned} P_1 : \text{Min } f(x) &= \frac{50}{x_1} + \frac{20}{x_2} + x_1 x_2 \\ \text{s.t. } &1 - x_1 \leq 0, \\ &1 - x_2 \leq 0, \end{aligned}$$

is not convex on F . After discussing the four possible cases for $I(x)$ (\emptyset , $\{1\}$, $\{2\}$, and $\{1, 2\}$), we find a unique candidate for $I(x) = \emptyset$: $\bar{x} = (5, 2)^T$, with $\nabla f(\bar{x}) = (-49, -19)^T$. The case $I(x) = \{1, 2\}$, corresponding to $x = (1, 1)^T$, is represented in Fig. 4.10. To prove that $F^* = \{\bar{x}\}$ it is enough to show that f is coercive. Indeed, given $x \in F$, one has

$$f(x) \leq \gamma \Rightarrow x_1 x_2 \leq \gamma \Rightarrow [1 \leq x_1 \leq \gamma \wedge 1 \leq x_2 \leq \gamma].$$

Observe that the relaxed problem obtained by eliminating the linear constraints is a geometric optimization problem already solved (see Exercise 3.14):

$$\begin{aligned} P_2 : \text{Min } f(x) &= \frac{50}{x_1} + \frac{20}{x_2} + x_1 x_2 \\ \text{s.t. } &x \in \mathbb{R}_{++}^2. \end{aligned}$$

Since the optimal solution of P_2 is $\bar{x} = (5, 2)^T$ and $\bar{x} \in F_1$, we conclude that $F_1^* = \{\bar{x}\}$.

4.3.2 Quadratic Optimization

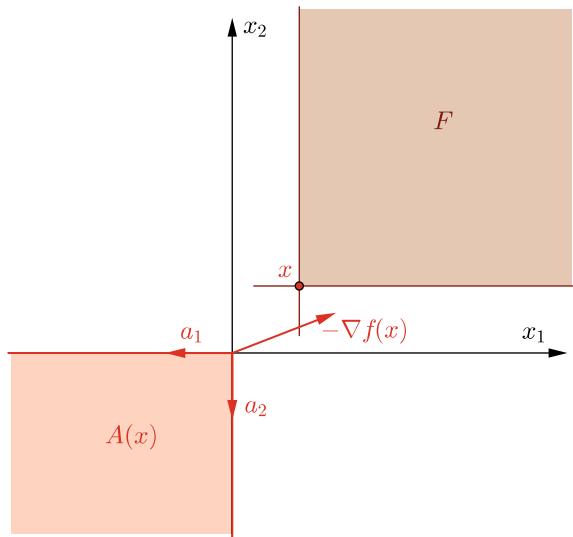
Linearly constrained convex quadratic optimization problems enjoy specific theoretical properties, as the so-called Frank–Wolfe Theorem [36], which guarantees the existence of optimal solutions even when the polyhedral feasible set is unbounded and the objective function is neither convex nor coercive.

Theorem 4.23 (Frank–Wolfe) *Let $Q \in \mathcal{S}_n$ and $c \in \mathbb{R}^n$, and suppose that*

$$\begin{aligned} P_Q : \text{Min } f(x) &= \frac{1}{2} x^T Q x - c^T x \\ \text{s.t. } &a_i^T x \leq b_i, \quad i \in I, \end{aligned} \tag{4.12}$$

is a bounded problem (not necessarily convex). Then, its optimal set F^ is nonempty.*

Fig. 4.10 The KKT conditions fail at $x = (1, 1)^T$



Proof The original proof given by Frank and Wolfe in [36] is beyond the scope of this book. An analytical direct proof was offered by Blum and Oettli in [14]. \square

The problem P_Q can be efficiently solved by ad hoc numerical methods which include active set, primal-dual, gradient projection, interior point, Frank–Wolfe, and simplex algorithms [12, 70, 87]. This type of problems directly arise in a variety of fields, e.g., in statistics, data fitting, portfolio problems (see [84], [87] and references therein), multiclass classification [84] (a widely used tool of machine learning), and also as subproblems in certain algorithms for nonlinear optimization problems (e.g., sequential quadratic and augmented Lagrangian algorithms). In [85], the reader can find several chapters related to quadratic optimization, e.g., a review on numerical methods (pp. 3–11) and applications to electrical engineering (pp. 27–36) and to chemical engineering (pp. 37–46).

Example 4.24 Consider the geometric problem consisting in the separation of two finite sets in \mathbb{R}^n , $\{u_1, \dots, u_p\}$ and $\{v_1, \dots, v_q\}$, by means of the thickest possible sandwich (the region of \mathbb{R}^n limited by two parallel hyperplanes, also called *strip* when $n = 2$); see Fig. 4.11.

A necessary and sufficient condition for the existence of such sandwiches is the existence of a hyperplane strictly separating the polytopes $U := \text{conv}\{u_1, \dots, u_p\}$ and $V := \text{conv}\{v_1, \dots, v_q\}$. It is easy to show that, if $U \cap V \neq \emptyset$, the optimal solution to this geometric problem can be obtained by computing a pair of vectors $(\bar{u}, \bar{v}) \in U \times V$ such that

$$\|\bar{u} - \bar{v}\| \leq \|u - v\|, \quad \forall (u, v) \in U \times V,$$

whose existence is consequence of the compactness of $U \times V$ and the continuity of the Euclidean norm. In fact, let $w := \bar{v} - \bar{u} \neq 0_n$. Since \bar{u} is the point of U closest

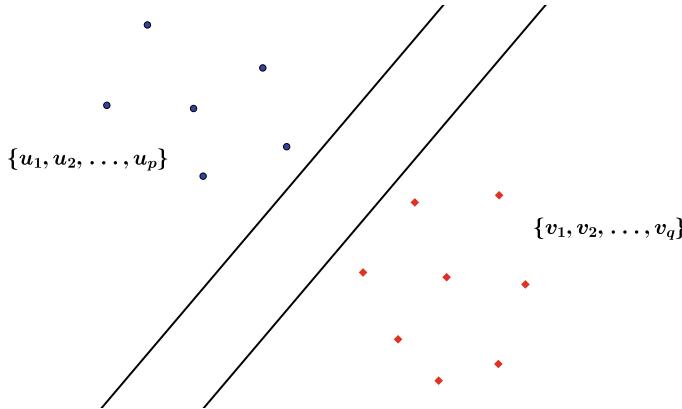


Fig. 4.11 Strip separating two finite sets

to $\bar{v} \neq \bar{u}$, $w^T(x - \bar{u}) \leq 0$ for all $x \in U$ (by the same argument as in the proof of Theorem 2.6). Similarly, $w^T(x - \bar{v}) \geq 0$ for all $x \in V$. So, the thickest sandwich separating U from V is

$$\{x \in \mathbb{R}^n : w^T \bar{v} \leq w^T x \leq w^T \bar{u}\},$$

whose *width* (also called *margin*) is $\|w\| = \|\bar{u} - \bar{v}\|$; see Fig. 4.12.

Since U and V are polytopes, they can be expressed (recall Example 2.14) as solution sets of the finite systems $\{c_i^T x \leq h_i, i \in I\}$ and $\{d_j^T x \leq e_j, j \in J\}$, respectively. Defining $y := \begin{pmatrix} u \\ v \end{pmatrix}$ and $A := [I_n \mid -I_n]$, and denoting by $0_{n \times n}$ the $n \times n$ null matrix, the optimization problem to be solved can be formulated as the quadratic optimization one

$$\begin{aligned} P_{U,V} : \text{Min}_{y \in \mathbb{R}^{2n}} \quad & y^T (A^T A) y \\ \text{s.t.} \quad & c_i^T [I_n \mid 0_{n \times n}] y \leq h_i, \quad i \in I, \\ & d_j^T [0_{n \times n} \mid I_n] y \leq e_j, \quad j \in J, \end{aligned}$$

which is convex, as the Gram matrix $A^T A$ is positive semidefinite (even though it is not positive definite because A is not column rank complete). So, if $\bar{y} = \begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix}$ is an optimal solution to $P_{U,V}$, the hyperplanes $w^T x = w^T \bar{u}$ and $w^T x = w^T \bar{v}$ support U at \bar{u} and V at \bar{v} , respectively.

In the simplest application of the *machine learning* methodology, the sets of vectors $\{u_1, \dots, u_p\}$ and $\{v_1, \dots, v_q\}$ in Example 4.24 represent two samples of a random vector, called *learning sets* (or *training sets*), of items which have been classified according to some criterion, e.g., in automatic diagnosis, patients sharing some symptom which are actually healthy or not (Classes I and II); In credit scoring, mortgage customers that respected or not its terms (Classes I and II), etc. In that case, if $\bar{y} = \begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix}$ is an optimal solution to $P_{U,V}$, its optimal value $\bar{y}^T (A^T A) \bar{y}$ is the

Fig. 4.12 Maximizing the margin between U and V

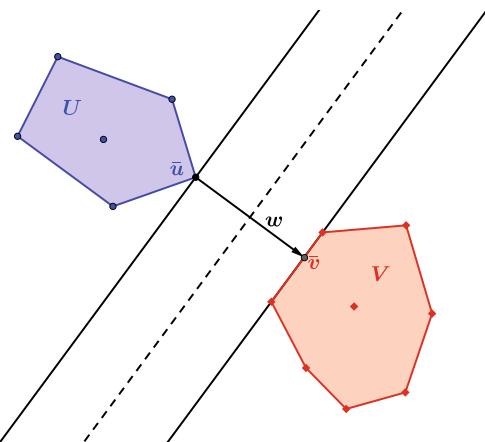
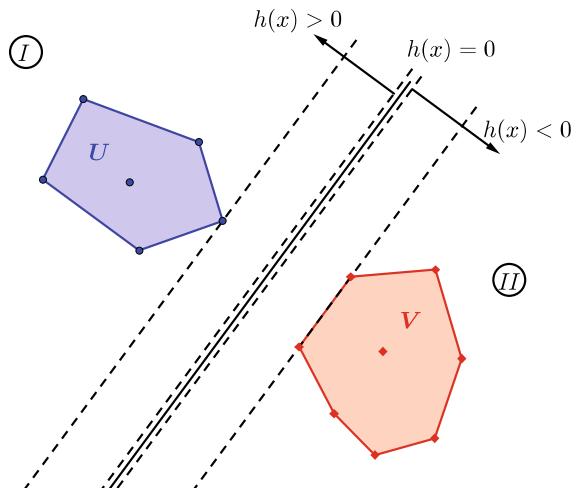


Fig. 4.13 Classification of new items



margin between the learning sets, and the affine function $h(x) := w^T x - w^T \left(\frac{\bar{v} + \bar{u}}{2}\right)$, with $w = \bar{v} - \bar{u}$, called *classifier*, allows to classify future items according to the sign of h on their corresponding observed vector x : As shown in Fig. 4.13, if $h(x) > 0$ (respectively, $h(x) < 0$), the item with observed vector x is likely in Class I (Class II, respectively), while the classification of future items with observed vector x such that $h(x) = 0$ (the hyperplane of symmetry of the thickest separating sandwich) is dubious.

The quadratic problem $P_{U,V}$ presents a double inconvenient: It may have multiple optimal solutions, and the computed optimal solution $\bar{y} = \left(\frac{\bar{u}}{\bar{v}}\right)$ may have many nonzero entries. Observe that, if $\bar{u}_k = \bar{v}_k = 0$, the k th component of the observations is useless for the classification and can be eliminated from the medical check or from the list of documents attached by the borrower to his/her application.

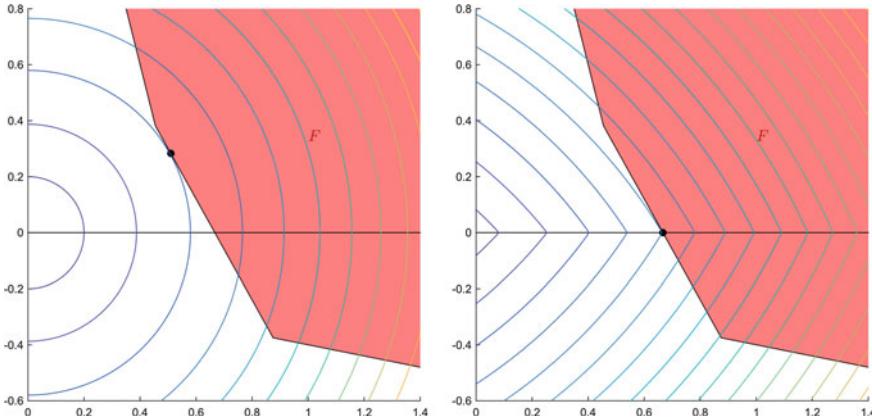


Fig. 4.14 Level curves and optimal solution to $\frac{1}{2}x^T x$ (left) and $\frac{1}{2}x^T x + \|x\|_1$ (right)

In order to overcome the first inconvenience, we can add to the objective function of $P_{U,V}$ a *regularization term* $\gamma \|y\|^2$, with $\gamma > 0$ selected by the decision maker, guaranteeing the existence of a unique optimal solution to the resulting convex quadratic optimization problem (thanks to the strong convexity of the regularization term). This optimal solution, depending on γ , approaches the optimal set of $P_{U,V}$ as γ decreases to 0.

Regarding the second inconvenience, we could follow a similar strategy, consisting in adding to the objective function of $P_{U,V}$ a *sparsity term* $\gamma s(y)$, with $s(y)$ denoting the number of nonzero components of y , and $\gamma > 0$. Unfortunately, the function s is not even continuous. For this reason, it is usually replaced by the ℓ_1 norm, as the optimal solutions of the resulting problems tend to be *sparse vectors* (i.e., vectors with many zero components).

For example, consider the quadratic optimization problem P_Q in (4.12) with $Q = I_2$ and $c = 0_2$, whose polyhedral feasible set F is represented in Fig. 4.14. The function s is equal to 0 at 0_2 , has a value of 1 at the axes, and is equal to 2 on the rest of the plane. In Fig. 4.14, we represent the optimal solution to P_Q and that of the modified problem P_Q^1 , which has the same feasible set but the objective function is equal to $\frac{1}{2}x^T x + \|x\|_1$. Hence, s is equal to 2 at the optimal solution of P_Q , while it has a value of 1 at the solution of P_Q^1 . The addition of the ℓ_1 norm to the objective function is responsible for the pointedness of the level curves of the resulting function. Therefore, in contrast to the strategy for the regularization term $\gamma \|\cdot\|^2$, the value of γ in the sparsity term should not be chosen too small.

We now show that the addition of a term $\gamma \|x\|_1$, with $\gamma > 0$, to the objective function of a convex quadratic problem

$$\begin{aligned} P_Q : \text{Min } & \frac{1}{2}x^T Qx - c^T x \\ \text{s.t. } & a_i^T x \leq b_i, i \in I, \end{aligned}$$

preserves these desirable properties. In fact, recalling the change of variables we have used at Subsubsection 1.1.5.6 to linearize ℓ_1 regression problems, we write $x = u - v$, with $u, v \in \mathbb{R}_+^n$, to reformulate the convex but nonquadratic problem

$$\begin{aligned} P_Q^1 : \text{Min } & \frac{1}{2}x^T Qx - c^T x + \gamma \|x\|_1 \\ \text{s.t. } & a_i^T x \leq b_i, i \in I, \end{aligned}$$

as the convex quadratic one

$$\begin{aligned} P_Q^2 : \text{Min } & \frac{1}{2}(u - v)^T Q(u - v) - c^T(u - v) + \gamma 1_n^T(u + v) \\ \text{s.t. } & a_i^T(u - v) \leq b_i, i \in I, \\ & u_i \geq 0, v_i \geq 0, i = 1, \dots, n. \end{aligned}$$

This modeling trick has been successfully used in different fields to get convex quadratic reformulations of certain optimality problems. For instance, in compressed sensing [35] one can find unconstrained optimization problems of the form

$$\text{Min}_{x \in \mathbb{R}^n} \frac{1}{2}x^T Qx - c^T x + \gamma \|x\|_1,$$

with $\gamma > 0$, which can be reformulated as a convex quadratic problem as above. The same change of variables reduces the convex parametric problem with convex quadratic objective

$$\begin{aligned} P^\gamma : \text{Min } & \frac{1}{2}x^T Qx - c^T x \\ \text{s.t. } & \|x\|_1 \leq \gamma, \end{aligned}$$

which arises in the *homotopy method* for variable selection [73] (where the optimal solutions to P^γ form, as the parameter γ increases, the so-called *path of minimizers*) into the linearly constrained quadratic problem

$$\begin{aligned} P_Q^\gamma : \text{Min } & \frac{1}{2}(u - v)^T Q(u - v) - c^T(u - v) \\ \text{s.t. } & 1_n^T(u + v) \leq \gamma. \\ & u_i \geq 0, v_i \geq 0, i = 1, \dots, n. \end{aligned}$$

An important variant to the least squares problem P_{LS} in (3.10) arising in machine learning is the so-called *ℓ_1 -penalized least absolute shrinkage and selection operator (LASSO)* [18],

$$P_{LASSO}^\gamma : \text{Min}_{x \in \mathbb{R}^n} f(x) = \frac{1}{2}\|Ax - b\|^2 + \gamma \|x\|_1,$$

where the parameter γ controls the weight of the penalization, whose main advantage over the LS estimator is the observed low density of its optimal solution, the so-called *LASSO estimator*. Due to the lack of differentiability of the penalizing term $\gamma \|x\|_1$, P_{LASSO}^γ is a hard problem for which specific first- and second-order methods have been proposed [18, Section 8]. However, as observed in [87], P_{LASSO}^γ can be more easily solved once it has been reformulated as a linearly constrained quadratic

optimization problem, by replacing the decision variable x by $u, v \in \mathbb{R}_+^n$ such that $x = u - v$.

Observe finally that the quadratic cone optimization problems of the type

$$\begin{aligned} P_K : \text{Min } & \frac{1}{2}x^T Qx - c^T x \\ \text{s.t. } & Ax + b \in K, \end{aligned}$$

where A is an $m \times n$ matrix, $b \in \mathbb{R}^m$ and K is a polyhedral convex cone in \mathbb{R}^m , can also be reduced to a linearly constrained quadratic optimization problem thanks to the Farkas Lemma 4.15. In fact, since K° is polyhedral too, we can write $K^\circ = \text{cone}\{y_1, \dots, y_p\}$. Since $K = K^{\circ\circ} = \{y_1, \dots, y_p\}^\circ$, we have that

$$Ax + b \in K \iff y_j^T(Ax + b) \leq 0, \quad \forall j \in \{1, \dots, p\}.$$

So, the conic constraint $Ax + b \in K$ can be replaced in P_K by the linear system $\{(y_j^T A)x \leq -y_j^T b, j = 1, \dots, p\}$.

The main reasons to devote a subsection to this particular class of convex optimization problems are its importance in practice and the fact that they can be solved with pen and paper when the number of inequality constraints is sufficiently small. Even more, as shown next, their optimal sets can be expressed by means of closed formulas in the favorable case where all constraints are linear equations, as it happens in the design of electric circuits (recall that the heat generated in a conductor is proportional to the resistance times the square of the intensity, so Q is diagonal and $c = 0_n$).

Proposition 4.25 (Quadratic optimization with linear inequality constraints) *Suppose that the problem P_Q in (4.12) is bounded and that $Q \in \mathcal{S}_n$ is positive semidefinite. Then, $F^* \neq \emptyset$ (in particular, $|F^*| = 1$ whenever Q is positive definite). Moreover, $\bar{x} \in F$ is an optimal solution to P_Q if and only if there exists $\bar{\lambda} \in \mathbb{R}^m$ such that:*

- (NC) $\bar{\lambda} \in \mathbb{R}_+^m$;
- (SC) $c - Q\bar{x} = \sum_{i \in I} \bar{\lambda}_i a_i$; and
- (CC) $\bar{\lambda}_i (b_i - a_i^T \bar{x}) = 0, \quad \forall i \in I$.

Proof It is a straightforward consequence of the Frank–Wolfe Theorem 4.23, Propositions 3.1 and 4.1, and Theorem 4.20. \square

The advantage of the quadratic optimization problems in comparison with those considered in Subsection 4.3.1 is that the KKT system to be solved for each possible value of $I(x)$,

$$\left\{ \begin{array}{l} c - Qx = \sum_{i \in I(x)} \lambda_i a_i \\ a_i^T x = b_i, \quad i \in I(x) \end{array} \right\}$$

is now linear.

A *multifunction* (or *set-valued mapping*) between two nonempty sets $X \subset \mathbb{R}^m$ and $Y \subset \mathbb{R}^n$ is a correspondence associating with each $x \in X$ a subset of Y .

Definition 4.26 The *metric projection* onto a closed set $\emptyset \neq C \subset \mathbb{R}^n$ is the multi-function $P_C : \mathbb{R}^n \rightrightarrows \mathbb{R}^n$ associating with each $y \in \mathbb{R}^n$ the set

$$P_C(y) = \{x \in C : d(y, x) = d(y, C)\},$$

that is, the set of global minima of the function $x \mapsto d(y, x) = \|x - y\|$ on C .

Obviously, $P_C(y) \neq \emptyset$ as $d(y, \cdot)$ is a continuous coercive function on \mathbb{R}^n and the set C is closed. For instance, if C is a sphere centered at z , $P_C(z) = C$ and $P_C(y) = C \cap \{z + \lambda(y - z) : \lambda \geq 0\}$ (a unique point) for all $y \neq z$.

Proposition 4.25 allows to compute $P_C(y)$ for any $y \in \mathbb{R}^n$, whenever C is a polyhedral convex set, but not to get an explicit expression for P_C .

Example 4.27 To compute the projection of $(3, 1, -1)^T$ onto the polyhedral cone K of Example 2.14,

$$K = \{x \in \mathbb{R}^3 : x_1 + x_2 - 2x_3 \leq 0, x_2 - x_3 \leq 0, -x_3 \leq 0\}$$

we must minimize $\|(x_1 - 3, x_2 - 1, x_3 + 1)^T\|^2$ on K ; i.e., we must solve the convex quadratic problem

$$\begin{aligned} P_Q : \text{Min } f(x) &= x^T x - (6, 2, -2)x \\ \text{s.t. } &x_1 + x_2 - 2x_3 \leq 0, \\ &x_2 - x_3 \leq 0, \\ &-x_3 \leq 0, \end{aligned}$$

which has a positive definite matrix $Q = 2I_3$, so that P_Q has a unique optimal solution. The KKT system for $I(x) = \{1\}$ is

$$\left\{ \begin{array}{l} -2x + (6, 2, -2)^T = \lambda_1(1, 1, -2)^T \\ x_1 + x_2 - 2x_3 = 0 \\ \lambda_1 \geq 0 \\ x_2 - x_3 < 0 \\ -x_3 < 0 \end{array} \right\},$$

whose unique solution is $(x^T, \lambda_1) = (2, 0, 1, 2)$. Thus, $P_K((3, 1, -1)^T) = \{(2, 0, 1)^T\}$, as shown in Fig. 4.15.

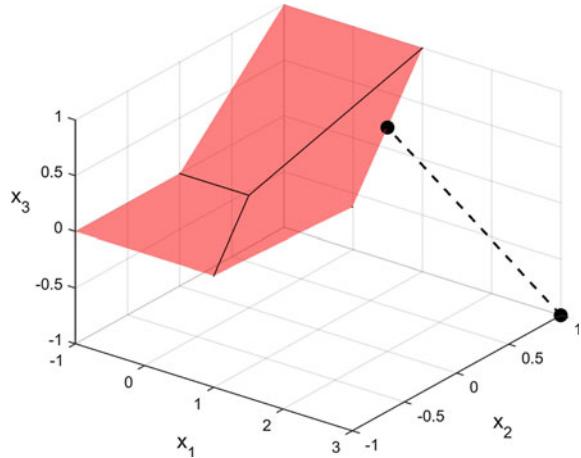
We now consider the particular case of the minimization of a quadratic function on an affine manifold.

Proposition 4.28 (Quadratic optimization with linear equality constraints) *Let*

$$\begin{aligned} P_Q : \text{Min } f(x) &= \frac{1}{2}x^T Qx - c^T x + b \\ \text{s.t. } &Mx = d, \end{aligned} \tag{4.13}$$

be such that $Q \in \mathcal{S}_n$ is positive definite, $c \in \mathbb{R}^n$, $b \in \mathbb{R}$, M is a full row rank $m \times n$

Fig. 4.15 Projection of $(3, 1, -1)^T$ onto the polyhedral cone K



matrix, and $d \in \mathbb{R}^m$. Then, the unique optimal solution to P_Q is

$$\bar{x} = Q^{-1}M^T(MQ^{-1}M^T)^{-1}(d - MQ^{-1}c) + Q^{-1}c. \quad (4.14)$$

Proof Let $a_i^T, i = 1, \dots, m$, be the rows of M . By assumption, we know that the set $\{a_i, i = 1, \dots, m\}$ is linearly independent.

Since Q is a positive definite symmetric matrix, f is coercive and strictly convex on \mathbb{R}^n by Proposition 3.1, and so on the affine manifold $F := \{x \in \mathbb{R}^n : Mx = d\}$. Therefore, we have that P has a unique optimal solution \bar{x} , which must satisfy the KKT conditions. Hence, by Proposition 4.25, there exists a multiplier vector $\lambda \in \mathbb{R}^m$ such that

$$Q\bar{x} - c + \sum_{i=1}^m \lambda_i a_i = Q\bar{x} - c + M^T\lambda = 0_n,$$

which, by the nonsingularity of Q , yields

$$\bar{x} = -Q^{-1}M^T\lambda + Q^{-1}c. \quad (4.15)$$

We now show that the symmetric matrix $MQ^{-1}M^T$ is positive definite. Indeed, given $y \in \mathbb{R}^m \setminus \{0_m\}$, one has

$$y^T(MQ^{-1}M^T)y = (M^Ty)^TQ^{-1}(M^Ty) > 0,$$

as Q^{-1} is positive definite and the columns of M^T are linearly independent. Thus, $MQ^{-1}M^T$ is nonsingular and (4.15) provides

$$d = M\bar{x} = -(MQ^{-1}M^T)\lambda + MQ^{-1}c.$$

This equation allows to obtain the expression of the multiplier vector corresponding to \bar{x} :

$$\lambda = -(MQ^{-1}M^T)^{-1}(d - MQ^{-1}c).$$

Replacing this vector λ in (4.15), and simplifying, one gets (4.14). \square

Example 4.29 Let

$$\begin{aligned} P : \text{Min } f(x) &= \frac{1}{2}x_1^2 + x_2^2 + \frac{3}{2}x_3^2 + x_1x_3 - 6x_1 - 4x_2 + 6 \\ \text{s.t. } x_1 + x_2 &= 4, \quad x_1 + 2x_3 = 3. \end{aligned}$$

We directly get the unique optimal solution to P from (4.14): $\bar{x} = \left(\frac{43}{11}, \frac{1}{11}, \frac{-5}{11}\right)^T$.

4.3.3 Some Closed Formulas

Proposition 4.28 has an immediate application to statistics, more concretely to ℓ_2 regression with interpolation: Given the point cloud $\{(t_i, s_i), i = 1, \dots, p\}$, with $t_i \neq t_j$ for all $i \neq j$, we try to get the polynomial $p(t) = x_1 + x_2t + \dots + x_{q+1}t^q$ of degree at most q which minimizes the Euclidean norm of the residual vector corresponding to the first m observations

$$r := \begin{pmatrix} x_1 + x_2t_1 + \dots + x_{q+1}t_1^q - s_1 \\ \vdots \\ x_1 + x_2t_m + \dots + x_{q+1}t_m^q - s_m \end{pmatrix} \in \mathbb{R}^m, \quad m < p,$$

and such that it interpolates the remaining points; i.e., it satisfies $p(t_i) = s_i$, $i = m+1, \dots, p$ (see Fig. 4.16).

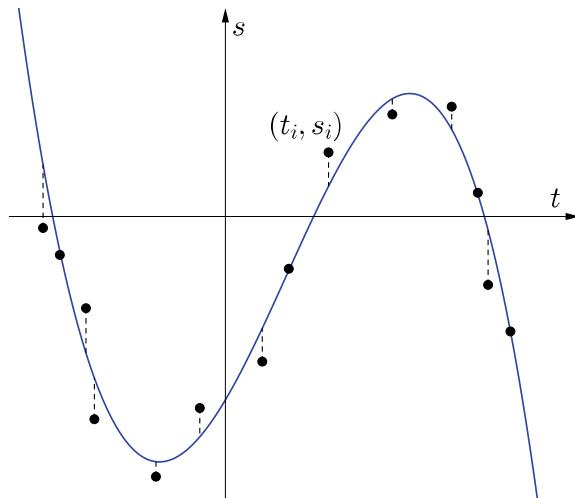
The problem to be solved is

$$\begin{aligned} P : \text{Min } f(x) &:= \sum_{i=1}^m \left(\sum_{j=1}^{q+1} x_j t_i^{j-1} - s_i \right)^2 \\ \text{s.t. } \sum_{j=1}^{q+1} x_j t_i^{j-1} &= s_i, \quad i = m+1, \dots, p, \end{aligned} \tag{4.16}$$

which is of the form of Proposition 4.28, with $Q = 2N^T N$, $c = 2N^T s$, and $b = \|s\|^2$, where

$$N := \begin{bmatrix} 1 & t_1 & \dots & t_1^q \\ 1 & t_2 & \dots & t_2^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_m & \dots & t_m^q \end{bmatrix} \quad \text{and} \quad s := \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_m \end{pmatrix},$$

Fig. 4.16 ℓ_2 regression with interpolation



while

$$M := \begin{bmatrix} 1 & t_{m+1} & \dots & t_{m+1}^q \\ 1 & t_{m+2} & \dots & t_{m+2}^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{q+1} & \dots & t_{q+1}^q \end{bmatrix} \quad \text{and} \quad d := \begin{pmatrix} s_{m+1} \\ s_{m+2} \\ \vdots \\ s_{q+1} \end{pmatrix}.$$

A straightforward application of Proposition 4.28 provides the following closed formula for this problem.

Corollary 4.30 (Regression with interpolation) *The unique optimal solution to the ℓ_2 regression problem with interpolation in (4.16) is the polynomial $\bar{p}(t) := \bar{x}_1 + \bar{x}_2 t + \dots + \bar{x}_{q+1} t^q$, where*

$$\bar{x} = (N^T N)^{-1} M^T \left(M(N^T N)^{-1} M^T \right)^{-1} \left(d - M(N^T N)^{-1} N^T s \right) + (N^T N)^{-1} N^T s.$$

We now apply Proposition 4.28 to the computation of $P_C(y)$, whenever C is an affine manifold.

Corollary 4.31 (Metric projection onto affine manifolds) *Let M be a full row rank $m \times n$ matrix. The metric projection of any $y \in \mathbb{R}^n$ onto the affine manifold $F = \{x \in \mathbb{R}^n : Mx = d\}$ is*

$$P_F(y) = \left\{ M^T (MM^T)^{-1} (d - My) + y \right\}, \quad (4.17)$$

and the distance from y to F is

$$d(y, F) = \left\| M^T (MM^T)^{-1} (d - My) \right\|.$$

Proof The problem to be solved is P_Q in (4.13), with

$$f(x) = \|x - y\|^2 = x^T x - 2y^T x + \|y\|^2.$$

Taking $Q = 2I_n$, $c = 2y$, and $b = \|y\|^2$ in (4.14), one gets (4.17). Moreover,

$$d(y, F) = \|P_F(y) - y\| = \left\| M^T (MM^T)^{-1} (d - My) \right\|,$$

which completes the proof. \square

In practice, it is convenient to compute $P_F(y)$ in two steps, without inverting the Gram matrix $G := MM^T$ of M , as follows:

- Step 1: Find w such that $\{Gw = d - My\}$.
- Step 2: Compute $P_F(y) = M^T w + y$.

The next result is an immediate consequence of Corollary 4.31.

Corollary 4.32 (The distance from the origin to a linear manifold) *Let M be a full row rank $m \times n$ matrix. The solution to the consistent system $\{Mx = d\}$ of minimum Euclidean norm is*

$$P_F(0_n) = \left\{ M^T (MM^T)^{-1} d \right\}. \quad (4.18)$$

Example 4.33 The point of the hyperplane $H = \{x \in \mathbb{R}^n : c^T x = d\}$, with $c \in \mathbb{R}^n \setminus \{0_n\}$ and $d \in \mathbb{R}$, closest to a given point y can be obtained by taking $M = c^T$ in (4.17), i.e.,

$$P_H(y) = \left\{ \|c\|^{-2} (d - c^T y) c + y \right\},$$

so the Euclidean distance from y to H is given by the well-known formula

$$d(y, H) = \|P_H(y) - y\| = \frac{|c^T y - d|}{\|c\|}. \quad (4.19)$$

Example 4.34 We are asked to compute the point of the affine manifold

$$F := \{x \in \mathbb{R}^3 : x_1 + x_2 + x_3 = 1, -x_1 - x_2 + x_3 = 0\}$$

closest to the origin (see Fig. 4.17). We have

$$M = \begin{bmatrix} 1 & 1 & 1 \\ -1 & -1 & 1 \end{bmatrix} \text{ and } d = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

We can apply (4.18) in two ways:

1. By matrix calculus:

$$P_F(0_3) = \{M^T G^{-1} d\} = \left\{ \begin{bmatrix} 1 & -1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \frac{3}{8} & \frac{1}{8} \\ \frac{1}{8} & \frac{1}{8} \end{bmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\} = \left\{ \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{2} \end{pmatrix} \right\}.$$

2. By Gauss elimination: The unique solution to $\{Gw = d\}$ is $w = \left(\frac{3}{8}, \frac{1}{8}\right)^T$. Thus,

$$P_F(0_3) = \{M^T w\} = \left\{ \begin{bmatrix} 1 & -1 \\ 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{pmatrix} \frac{3}{8} \\ \frac{1}{8} \end{pmatrix} \right\} = \left\{ \begin{pmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{2} \end{pmatrix} \right\}.$$

4.4 Arbitrarily Constrained Convex Optimization*

The rest of the chapter is devoted to the study of convex optimization problems of the form

$$\begin{aligned} P : \text{Min } f(x) \\ \text{s.t. } g_i(x) \leq 0, \quad i \in I, \\ x \in C, \end{aligned} \tag{4.20}$$

where $\emptyset \neq C \subset \mathbb{R}^n$ is a convex set, the functions g_i are convex on C , $i \in I$, and f is convex on the feasible set F of P . For this type of problems, we are not able to give a closed solution, but we can at least obtain optimal solutions in an analytic way via KKT conditions. When P arises in industrial production planning, any perturbation of the right-hand side of $g_i(x) \leq 0$ represents a variation of the available amount of the i th recourse. Sensitivity analysis allows to estimate the variation of the optimal value under this type of perturbations when there exists a so-called sensitivity vector which is related to the KKT necessary conditions at an optimal solution to P . Stopping rules for the numerical optimization algorithms for P can be obtained from either the mentioned KKT necessary conditions or by solving simultaneously P and a suitable dual problem providing lower bounds for the optimal value of P .

4.4.1 Sensitivity Analysis

This subsection deals with the *parametric problem* that results from replacing 0 by a parameter $z_i \in \mathbb{R}$ in the functional constraint $g_i(x) \leq 0$, $i \in I = \{1, \dots, m\}$:

$$\begin{aligned} P(z) : \text{Min } f(x) \\ \text{s.t. } g_i(x) \leq z_i, \quad i \in I, \\ x \in C. \end{aligned}$$

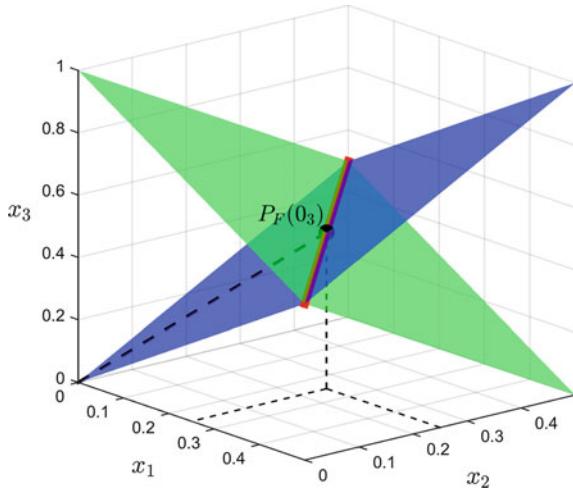


Fig. 4.17 Minimum Euclidean norm solution to the system $\{x_1+x_2+x_3=1, -x_1-x_2+x_3=0\}$

The parameter vector $z = (z_1, \dots, z_m)^T \in \mathbb{R}^m$ will be called *right-hand side vector* in $P(z)$. The feasible set and the optimal value of $P(z)$ will be denoted by $\mathcal{F}(z)$ and $\vartheta(z)$, respectively, with the convention that $\vartheta(z) = +\infty$ when $\mathcal{F}(z) = \emptyset$. Obviously, $P(0_m) \equiv P$ and $\vartheta(0_m) = v(P) \in \overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ (the *extended real line*). Observe that $\mathcal{F} : \mathbb{R}^m \rightrightarrows \mathbb{R}^n$ is a multifunction, while $\vartheta : \mathbb{R}^m \rightarrow \overline{\mathbb{R}}$ is an *extended real function*, called *feasible set multifunction* and *value function*, respectively. Both the *value function* ϑ and the *feasible set multifunction* \mathcal{F} are represented through their corresponding graphs

$$\text{gph } \vartheta := \left\{ \begin{pmatrix} z \\ y \end{pmatrix} \in \mathbb{R}^{m+1} : y = \vartheta(z) \right\}$$

and

$$\text{gph } \mathcal{F} := \left\{ \begin{pmatrix} z \\ x \end{pmatrix} \in \mathbb{R}^{m+n} : x \in \mathcal{F}(z) \right\}.$$

The *epigraph* of ϑ is the set

$$\text{epi } \vartheta := \left\{ \begin{pmatrix} z \\ y \end{pmatrix} \in \mathbb{R}^{m+1} : \vartheta(z) \leq y \right\}.$$

The *domains* of \mathcal{F} and ϑ are

$$\text{dom } \mathcal{F} := \{z \in \mathbb{R}^m : \mathcal{F}(z) \neq \emptyset\} \quad \text{and} \quad \text{dom } \vartheta := \{z \in \mathbb{R}^m : \vartheta(z) < +\infty\},$$

which obviously coincide. Observe that $z \in \text{int dom } \vartheta$ when small perturbations of z preserve the consistency of $P(z)$.

We now study the properties of ϑ , whose argument z is interpreted as a perturbation of 0_m . One of these properties is the convexity that we define now for extended functions. To handle these functions, we must first extend the algebraic operations (sum and product) and the natural ordering of \mathbb{R} to the extended real line $\overline{\mathbb{R}}$.

Concerning the sum, inspired by the algebra of limits of sequences, we define

$$\alpha + (+\infty) = (+\infty) + \alpha = +\infty, \quad \forall \alpha \in \mathbb{R} \cup \{+\infty\}$$

and

$$\alpha + (-\infty) = (-\infty) + \alpha = -\infty, \quad \forall \alpha \in \mathbb{R} \cup \{-\infty\}.$$

In the indeterminate case, it is convenient to define

$$(-\infty) + (+\infty) = (+\infty) + (-\infty) = +\infty. \quad (4.21)$$

In this way, the sum is well defined on $\overline{\mathbb{R}}$ and is commutative.

The product of elements of $\overline{\mathbb{R}}$ is also defined as in the algebra of limits (e.g., $\alpha(+\infty) = +\infty$ if $\alpha > 0$), with the following convention for the indeterminate cases:

$$0(+\infty) = (+\infty)0 = 0(-\infty) = (-\infty)0 = 0. \quad (4.22)$$

So, the product is also well defined on $\overline{\mathbb{R}}$ and is commutative (but $(\overline{\mathbb{R}}; +, \cdot)$ is not a field as $+\infty$ and $-\infty$ do not have inverse elements).

The extension of the ordering of \mathbb{R} to $\overline{\mathbb{R}}$ is made by the convention that

$$-\infty < \alpha < +\infty, \quad \forall \alpha \in \mathbb{R}.$$

The extension of the absolute value from \mathbb{R} to $\overline{\mathbb{R}}$ consists in defining $|+\infty| = |-\infty| = +\infty$, and has similar properties. The calculus rules for sums, products, and inequalities on $\overline{\mathbb{R}}$ are similar to those of \mathbb{R} , just taking into account (4.21) and (4.22). For instance, if $\alpha, \beta \in \{\pm\infty\}$, then $|\alpha - \beta| = +\infty$. Any nonempty subset of $\overline{\mathbb{R}}$ (defined as in \mathbb{R}) has an infimum (supremum), possibly $-\infty$ ($+\infty$, respectively). We also define $\inf \emptyset = +\infty$ and $\sup \emptyset = -\infty$.

We are now in a position to define convexity of extended real functions, which, unlike Definition 2.26, does not involve a particular convex set C of \mathbb{R}^n .

Definition 4.35 The extended real function $h : \mathbb{R}^n \rightarrow \overline{\mathbb{R}}$ is *convex* if

$$h((1 - \mu)x + \mu y) \leq (1 - \mu)h(x) + \mu h(y), \quad \forall x, y \in \mathbb{R}^n, \quad \forall \mu \in]0, 1[,$$

and it is *concave* when $-h$ is convex.

It is easy to prove that h is convex if and only if $\text{epi } h$ is a convex subset of \mathbb{R}^{n+1} . Hence, the supremum of convex functions (in particular the supremum of affine functions) is also convex. Since linear mappings between linear spaces preserve

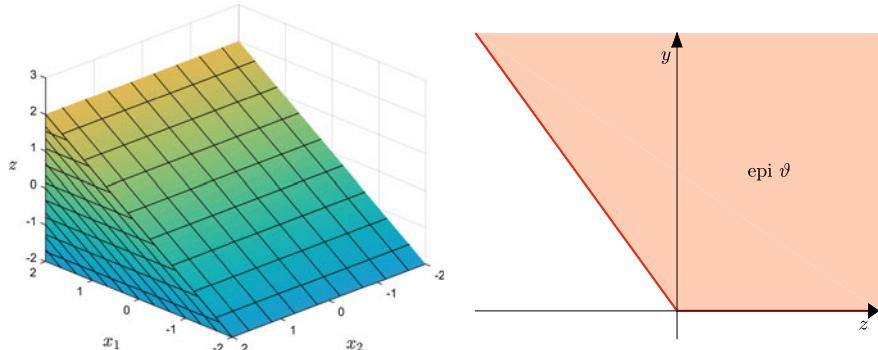


Fig. 4.18 $\text{gph } \mathcal{F}$ (left) and $\text{gph } \vartheta$ (right)

the convexity of sets and $\text{dom } h$ is the image of $\text{epi } h$ by the vertical projection $\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^{n+1} \mapsto x \in \mathbb{R}^n$, if h is convex then $\text{dom } h$ is convex too.

Definition 4.36 A given point $\hat{x} \in C$ is a *Slater point* for P if $g_i(\hat{x}) < 0$ for all $i \in I$ (that is, $\hat{x} \in F$ and $I(\hat{x}) = \emptyset$). We say that the problem P satisfies the *Slater constraint qualification (SCQ)* when there exists a Slater point.

The next example shows that ϑ might be nondifferentiable at 0_m even when SCQ holds, although it is a convex function, as we shall see in Theorem 4.38. It can even be noncontinuous; see Exercise 4.16.

Example 4.37 Consider the convex optimization problem

$$\begin{aligned} P : \text{Min } f(x) &= |x_1| + x_2 \\ \text{s.t. } g(x) &= x_1 \leq 0 \\ x &\in C = \mathbb{R} \times \mathbb{R}_+. \end{aligned}$$

It can be easily checked that P satisfies SCQ, $F^* = \{0_2\}$, $v(P) = 0$, $\mathcal{F}(z) =]-\infty, z] \times \mathbb{R}_+$, $\vartheta(z) = -z$, if $z < 0$, and $\vartheta(z) = 0$, when $z \geq 0$. In Fig. 4.18, we show $\text{gph } \mathcal{F} = \{(z, x_1, x_2)^T \in \mathbb{R}^3 : x_1 \geq z\}$, $\text{gph } \vartheta = \mathbb{R}_+ \begin{pmatrix} -1 \\ 1 \end{pmatrix} \cup (\mathbb{R} \times \{0\})$, and $\text{epi } \vartheta = \text{cone}\{\begin{pmatrix} -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}\}$. Finally, observe that ϑ is continuous on \mathbb{R} and differentiable on $\mathbb{R} \setminus \{0\}$.

Theorem 4.38 (Convexity of the value function) *The value function ϑ of a convex optimization problem P is convex. Moreover, if P satisfies SCQ, then $0_m \in \text{int dom } \vartheta$.*

Proof Observe that the restriction of ϑ to $\text{dom } \vartheta$ takes values in $\mathbb{R} \cup \{-\infty\}$. Let $y, z \in \mathbb{R}^n$ and $\mu \in]0, 1[$. If either $\vartheta(x) = +\infty$ or $\vartheta(y) = +\infty$, then $(1 - \mu)\vartheta(x) + \mu\vartheta(y) = +\infty$, according to the calculus rules on $\overline{\mathbb{R}}$. Hence, we may assume that $y, z \in \text{dom } \vartheta$.

We claim the inclusion $A \subset B$ for the sets A and B defined as follows:

$$A := \{(1 - \mu)x_1 + \mu x_2 : x_1, x_2 \in C, \\ (1 - \mu)g_i(x_1) + \mu g_i(x_2) \leq (1 - \mu)y_i + \mu z_i, \forall i \in I\}$$

and

$$B := \{x \in C : g_i(x) \leq (1 - \mu)y_i + \mu z_i, \forall i \in I\} = \mathcal{F}((1 - \mu)y + \mu z).$$

Indeed, take an arbitrary $x := (1 - \mu)x_1 + \mu x_2 \in A$. By the convexity of g_i on C , $i \in I$, one has

$$g_i(x) \leq (1 - \mu)g_i(x_1) + \mu g_i(x_2) \leq (1 - \mu)y_i + \mu z_i, \quad \forall i \in I,$$

so we have that $x \in B$.

The inclusion $A \subset B$ is preserved when one takes images by f , i.e., $f(A) \subset f(B)$, and the order is inverted when one takes infima: $\inf f(B) \leq \inf f(A)$. By the definition of ϑ and the convexity of f , we get

$$\begin{aligned} \vartheta((1 - \mu)y + \mu z) &= \inf\{f(x) : x \in \mathcal{F}((1 - \mu)y + \mu z)\} \\ &= \inf f(B) \leq \inf f(A) \\ &= \inf \{f((1 - \mu)x_1 + \mu x_2) : x_1, x_2 \in C, \\ &\quad (1 - \mu)g_i(x_1) + \mu g_i(x_2) \leq (1 - \mu)y_i + \mu z_i \forall i \in I\} \quad (4.23) \\ &\leq \inf \{(1 - \mu)f(x_1) + \mu f(x_2) : x_1, x_2 \in C, \\ &\quad (1 - \mu)g_i(x_1) + \mu g_i(x_2) \leq (1 - \mu)y_i + \mu z_i \forall i \in I\}. \end{aligned}$$

Consider now the sets

$$D := \{(x_1, x_2) \in C \times C : g_i(x_1) \leq y_i \text{ and } g_i(x_2) \leq z_i, \forall i \in I\}$$

and

$$E := \{(x_1, x_2) \in C \times C : (1 - \mu)g_i(x_1) + \mu g_i(x_2) \leq (1 - \mu)y_i + \mu z_i, \forall i \in I\}.$$

Since, for any $i \in I$,

$$[g_i(x_1) \leq y_i \text{ and } g_i(x_2) \leq z_i] \Rightarrow (1 - \mu)g_i(x_1) + \mu g_i(x_2) \leq (1 - \mu)y_i + \mu z_i,$$

we have $D \subset E$. By (4.23), the latter inclusion, and the fact that the infimum of the sum of two sets of real numbers is the sum of their respective infima, one has

$$\begin{aligned}
\vartheta((1-\mu)y + \mu z) &\leq \inf\{(1-\mu)f(x_1) + \mu f(x_2) : (x_1, x_2) \in E\} \\
&\leq \inf\{(1-\mu)f(x_1) + \mu f(x_2) : (x_1, x_2) \in D\} \\
&= \inf\{(1-\mu)f(x_1) : x_1 \in C, g_i(x_1) \leq y_i, \forall i \in I\} \\
&\quad + \inf\{\mu f(x_2) : x_2 \in C, g_i(x_2) \leq z_i, \forall i \in I\} \\
&= (1-\mu)\vartheta(y) + \mu\vartheta(z).
\end{aligned}$$

Hence, ϑ is convex.

Assume now that P satisfies SCQ. Let $\hat{x} \in C$ be such that $g_i(\hat{x}) < 0$ for all $i \in I$. Let $\rho := \min_{i \in I}(-g_i(\hat{x})) > 0$, that is, $\max_{i \in I} g_i(\hat{x}) = -\rho$.

Given $z \in \rho\mathbb{B}$, one has $|z_i| \leq \rho$, so $g_i(\hat{x}) \leq -\rho \leq z_i, i \in I$. Thus, $\hat{x} \in \mathcal{F}(z)$ and one has $z \in \text{dom } \mathcal{F} = \text{dom } \vartheta$. Since $\rho\mathbb{B} \subset \text{dom } \vartheta$, we conclude that $0_m \in \text{int dom } \vartheta$, which completes the proof. \square

In Exercise 4.16, ϑ is finite valued on $\text{dom } \vartheta$. We now show that this fact is a consequence of the convexity of ϑ and its finiteness at some point of $\text{int dom } \vartheta = \mathbb{R}$.

Corollary 4.39 *If $0_m \in \text{int dom } \vartheta$ and $\vartheta(0_m) \in \mathbb{R}$, then ϑ is finite on the whole of $\text{dom } \vartheta$.*

Proof Suppose that $0_m \in \text{int dom } \vartheta$, and assume, by contradiction, that there exists a point $z^1 \in \text{dom } \vartheta$ such that $\vartheta(z^1) = -\infty$. Let $\varepsilon > 0$ be such that $\varepsilon\mathbb{B} \subset \text{dom } \vartheta$. The point $z^2 := -\frac{\varepsilon}{\|z^1\|}z^1 \in \varepsilon\mathbb{B}$ satisfies $0_m \in]z^1, z^2[$. Let $\mu \in]0, 1[$ be such that $0_m = (1-\mu)z^1 + \mu z^2$. Then, due to the convexity of ϑ and the fact that $\vartheta(z^2) \in \mathbb{R} \cup \{-\infty\}$, we have

$$\vartheta(0_m) \leq (1-\mu)\vartheta(z^1) + \mu\vartheta(z^2) = -\infty,$$

so that $\vartheta(0_m) = -\infty$ (contradiction). \square

As a consequence of Theorem 4.38, $\text{epi } \vartheta$ is convex, but it can be closed (as in Example 4.37) or not. It always satisfies the inclusion $\text{gph } \vartheta \subset \text{bd cl epi } \vartheta$ because $\text{gph } \vartheta \subset \text{epi } \vartheta$ and, for any $x \in \text{dom } \vartheta$, $(x, \vartheta(x) - \frac{1}{k})^T \notin \text{epi } \vartheta$ for all $k \in \mathbb{N}$. Thus, by the supporting hyperplane theorem, there exists a hyperplane supporting $\text{cl epi } \vartheta$ at any point of $\text{gph } \vartheta$. This hyperplane might be unique or not, and when it is unique, it can be vertical or not. For instance, in Example 4.37, any line of the form $y = -\lambda z$ with $\lambda \in [0, 1]$ supports $\text{epi } \vartheta$ at 0_2 , while the vertical line $\text{span}\left\{\begin{pmatrix} 0 \\ 1 \end{pmatrix}\right\}$ does not support $\text{epi } \vartheta$.

Theorem 4.40 (Sensitivity theorem) *If P is bounded and satisfies SCQ, then ϑ is finite valued on $\text{dom } \vartheta$ and there exists $\lambda \in \mathbb{R}_+^m$ such that*

$$\vartheta(z) \geq \vartheta(0_m) - \lambda^T z, \quad \forall z \in \text{dom } \vartheta. \tag{4.24}$$

If, additionally, ϑ is differentiable at 0_m , then the unique $\lambda \in \mathbb{R}_+^m$ satisfying (4.24) is $-\nabla \vartheta(0_m)$.

Proof By Theorem 4.38, ϑ is convex and $0_m \in \text{int dom } \vartheta$. Since $\vartheta(0_m) = v(P) \in \mathbb{R}$, Corollary 4.39 allows to assert that ϑ is finite valued on $\text{dom } \vartheta$.

Since ϑ is convex and $0_m \in \text{int dom } \vartheta$, by Proposition 2.33, there exists a non-vertical hyperplane $a^T x + a_{n+1}x_{n+1} = b$, with $a_{n+1} \neq 0$, which supports $\text{cl epi } \vartheta$ at $(\frac{0_m}{\vartheta(0_m)})$. Then, $a_{n+1}\vartheta(0_m) = b$, and we can assume that

$$a^T z + a_{n+1}z_{n+1} \leq b, \quad \forall \left(\begin{array}{c} z \\ z_{n+1} \end{array} \right) \in \text{cl epi } \vartheta. \quad (4.25)$$

Since $(\frac{0_m}{\vartheta(0_m)+\delta}) \in \text{epi } \vartheta$ for all $\delta \geq 0$, we have $a_{n+1} < 0$. Dividing by $|a_{n+1}|$ both members of (4.25), and defining $\gamma := \frac{a}{|a_{n+1}|}$ and $\beta := \frac{b}{|a_{n+1}|}$, one gets

$$\gamma^T z - z_{n+1} \leq \beta, \quad \forall \left(\begin{array}{c} z \\ z_{n+1} \end{array} \right) \in \text{gph } \vartheta,$$

with $\beta = \gamma^T 0_m - \vartheta(0_m) = -\vartheta(0_m)$. In other words,

$$\text{gph } \vartheta \subset \left\{ \left(\begin{array}{c} z \\ z_{n+1} \end{array} \right) \in \mathbb{R}^{m+1} : \left(\begin{array}{c} \gamma \\ -1 \end{array} \right)^T \left[\left(\begin{array}{c} z \\ z_{n+1} \end{array} \right) - \left(\begin{array}{c} 0_m \\ \vartheta(0_m) \end{array} \right) \right] \leq 0 \right\} \quad (4.26)$$

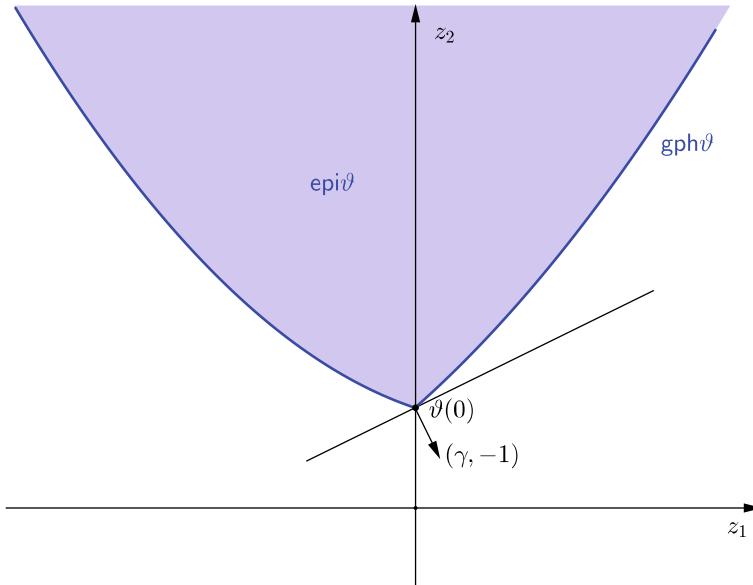


Fig. 4.19 Geometric interpretation of the subgradients when $m = 1$

(the vectors $\gamma \in \mathbb{R}^m$ satisfying (4.26) are called *subgradients* of ϑ at 0_m ; see Fig. 4.19.)

Given $z \in \text{dom } \vartheta$, as $\begin{pmatrix} z \\ \vartheta(z) \end{pmatrix} \in \text{gph } \vartheta$, we have

$$\begin{pmatrix} \gamma \\ -1 \end{pmatrix}^T \begin{pmatrix} z \\ \vartheta(z) - \vartheta(0_m) \end{pmatrix} = \gamma^T z - \vartheta(z) + \vartheta(0_m) \leq 0.$$

Hence, the vector $\lambda := -\gamma$ satisfies (4.24).

We now prove that $\lambda \in \mathbb{R}_+^m$ by contradiction. Assume that λ has a negative component $\lambda_i < 0$. Since $0_m \in \text{int dom } \vartheta$, there exists a sufficiently small positive number ε such that $\varepsilon \mathbb{B} \subset \text{dom } \vartheta$. Then, $\varepsilon e_i \in \text{dom } \vartheta$ and, by (4.24), one has

$$\vartheta(\varepsilon e_i) \geq \vartheta(0_m) - \lambda^T(\varepsilon e_i) = v(P) - \varepsilon \lambda_i > v(P). \quad (4.27)$$

But $\varepsilon e_i \in \mathbb{R}_+^m$ implies $\mathcal{F}(0_m) \subset \mathcal{F}(\varepsilon e_i)$, which in turn implies $v(P) \geq \vartheta(\varepsilon e_i)$, in contradiction with (4.27).

Finally, the case where ϑ is differentiable at 0_m is a direct consequence of the last assertion in Proposition 2.33. \square

A vector $\lambda \in \mathbb{R}_+^m$ satisfying (4.24) is said to be a *sensitivity vector* for P . The geometrical meaning of (4.24) is that there exists an affine function whose graph contains $\begin{pmatrix} 0_m \\ \vartheta(0_m) \end{pmatrix}$ and is a lower approximation of ϑ .

In Example 4.37, the sensitivity vectors (scalars here as $m = 1$) for P are the elements of the interval $[0, 1]$; see Fig. 4.18.

When ϑ is differentiable at 0_m , then the unique sensitivity vector is $\lambda = -\nabla \vartheta(0_m)$, which will be interpreted in Chapter 5 as the steepest descent direction for ϑ at 0_m (the search direction used by the steepest descent method to improve the current iterate). Alternatively, if ϑ is not differentiable at 0_m and $t > 0$ is sufficiently small, we have the following lower bound for the variation of ϑ in the direction of e_i , $i = 1, \dots, m$:

$$\frac{\vartheta(te_i) - \vartheta(0_m)}{t} \geq -\lambda_i.$$

As the section functions of ϑ have side derivatives, we can write

$$\vartheta'(0_m; e_i) = \lim_{t \searrow 0} \frac{\vartheta(te_i) - \vartheta(0_m)}{t} \geq -\lambda_i,$$

that is, the rate of variation of ϑ in the direction e_i is at least $-\lambda_i$.

In Example 4.37, $\vartheta'(0; 1) \geq -\lambda$ for all $\lambda \in [0, 1]$, so $\vartheta'(0; 1) \geq 0$. Similarly, $\vartheta'(0; -1) \geq \lambda$ for all $\lambda \in [0, 1]$, so $\vartheta'(0; -1) \geq 1$. Actually, one has that $\vartheta'(0; 1) = 0$ and $\vartheta'(0; -1) = 1$.

4.4.2 Optimality Conditions

Given a feasible solution \tilde{x} of the convex optimization problem P in (4.20) and a vector $\lambda \in \mathbb{R}_+^m$, we have

$$\inf \left\{ f(x) + \sum_{i \in I} \lambda_i g_i(x) : x \in C \right\} \leq f(\tilde{x}) + \sum_{i \in I} \lambda_i g_i(\tilde{x}) \leq f(\tilde{x}),$$

so one has that

$$h(\lambda) := \inf \left\{ f(x) + \sum_{i \in I} \lambda_i g_i(x) : x \in C \right\} \leq v(P).$$

Thus, $h(\lambda)$ is a lower bound for $v(P)$. In other words, the maximization of $h(\lambda)$ for $\lambda \in \mathbb{R}_+^m$ provides a dual problem for P . This motivates the study of the function depending on x and λ involved in the definition of h .

Definition 4.41 The *Lagrange function* of problem P in (4.20) is $L : C \times \mathbb{R}_+^m \rightarrow \mathbb{R}$ given by

$$L(x, \lambda) = f(x) + \sum_{i \in I} \lambda_i g_i(x).$$

We now prove that, if we know a sensitivity vector, we can reformulate P as an unconstrained convex problem.

Theorem 4.42 (Reduction to an unconstrained problem) *If λ is a sensitivity vector for a bounded convex problem P , then*

$$v(P) = \inf_{x \in C} L(x, \lambda).$$

Proof We associate with a given $x \in C$ the vector $z := (g_1(x), \dots, g_m(x))^T$. Since $x \in \mathcal{F}(z)$, we have $z \in \text{dom } \vartheta$ and $f(x) \geq \vartheta(z)$. Moreover, by the assumption on λ , we get

$$f(x) \geq \vartheta(z) \geq v(P) - \lambda^T z = v(P) - \sum_{i \in I} \lambda_i g_i(x),$$

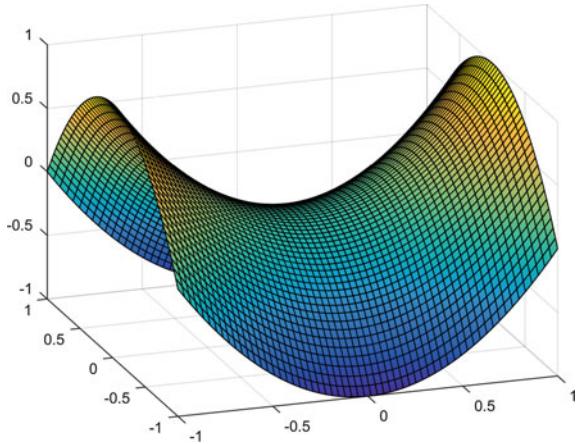
which yields $L(x, \lambda) \geq v(P)$. Taking the infimum of both sides for $x \in C$, we deduce $v(P) \leq \inf_{x \in C} L(x, \lambda)$.

Conversely, as $\lambda \in \mathbb{R}_+^m$, one has

$$\begin{aligned} \inf_{x \in C} L(x, \lambda) &\leq \inf \{L(x, \lambda) : x \in C, g_i(x) \leq 0, \forall i \in I\} \\ &\leq \inf \{f(x) : x \in C, g_i(x) \leq 0, \forall i \in I\} = v(P). \end{aligned}$$

We thus conclude that $v(P) = \inf_{x \in C} L(x, \lambda)$. □

Fig. 4.20 $(0, 0)$ is a saddle point of $(x, \lambda) \mapsto x^2 - \lambda^2$ on \mathbb{R}^2



The Lagrange function for the problem in Example 4.37 is $L(x, \lambda) = |x_1| + x_2 + \lambda x_1$. Taking the sensitivity vector (here a scalar) $\lambda = 1$, one has, for any $x \in C = \mathbb{R} \times \mathbb{R}_+$, $L(x, 1) = x_2$, if $x_1 < 0$, and $L(x, 1) = 2x_1 + x_2$, if $x_1 \geq 0$, so we have that $\inf_{x \in C} L(x, 1) = 0 = v(P)$.

Theorem 4.43 (Saddle point theorem) *Assume that P is a bounded convex problem satisfying SCQ. Let $\bar{x} \in C$. Then, $\bar{x} \in F^*$ if and only if there exists $\bar{\lambda} \in \mathbb{R}_+^m$ such that:*

- (NC) $\bar{\lambda} \in \mathbb{R}_+^m$;
- (SPC) $L(\bar{x}, \lambda) \leq L(\bar{x}, \bar{\lambda}) \leq L(x, \bar{\lambda}), \quad \forall x \in C, \lambda \in \mathbb{R}_+^m$; and
- (CC) $\bar{\lambda}_i g_i(\bar{x}) = 0, \quad \forall i \in I$.

The new acronym (SPC) refers to *saddle point condition*.

Remark 4.44 (Comment previous to the proof). (SPC) can be interpreted by observing that $(\bar{x}, \bar{\lambda})$ is a *saddle point* for the function $(x, \lambda) \mapsto L(x, \lambda)$ on $C \times \mathbb{R}_+^m$ as \bar{x} is a minimum of $L(x, \bar{\lambda})$ on C , while $\bar{\lambda}$ is a maximum of $L(\bar{x}, \lambda)$ on \mathbb{R}_+^m . For instance, 0_2 is a saddle point for the function $(x, \lambda) \mapsto x^2 - \lambda^2$ on \mathbb{R}^2 , as Fig. 4.20 shows. One can easily check that $(0, 0, 1)$ is a saddle point for the problem in Example 4.37, as $L(x, 1) = |x_1| + x_2 + x_1 \geq 0 = L(0, 0, 1)$ for all $x \in C$ and $L(0, 0, \lambda) = 0 = L(0, 0, 1)$ for all $\lambda \in \mathbb{R}_+$.

Proof Throughout this proof, we represent by (SPC1) and (SPC2) the first and the second inequalities in (SPC), respectively.

Let $\bar{x} \in F^*$. Under the assumptions on P , Theorem 4.40 implies the existence of a sensitivity vector $\bar{\lambda} \in \mathbb{R}_+^m$ for P . Obviously, $\bar{\lambda}$ satisfies (NC). Then, by Theorem 4.42, we have that $v(P) = \inf_{x \in C} L(x, \bar{\lambda})$. Since $\bar{x} \in F^* \subset F$ and $\bar{\lambda} \in \mathbb{R}_+^m$, we obtain

$$f(\bar{x}) = v(P) = \inf_{x \in C} L(x, \bar{\lambda}) \leq L(\bar{x}, \bar{\lambda}) = f(\bar{x}) + \sum_{i \in I} \bar{\lambda}_i g_i(\bar{x}) \leq f(\bar{x}). \quad (4.28)$$

From (4.28), we get, on the one hand, that $\sum_{i \in I} \bar{\lambda}_i g_i(\bar{x}) = 0$, i.e., $\bar{\lambda}_i g_i(\bar{x}) = 0$ for $i \in I$, which amounts to say that P satisfies (CC). On the other hand, since $L(\bar{x}, \bar{\lambda}) = \inf_{x \in C} L(x, \bar{\lambda})$, we have $L(\bar{x}, \bar{\lambda}) \leq L(x, \bar{\lambda})$ for all $x \in C$, so (SPC2) holds. In order to prove that (SPC1) also holds, take an arbitrary $\lambda \in \mathbb{R}_+^m$. By (CC),

$$L(\bar{x}, \bar{\lambda}) - L(\bar{x}, \lambda) = \sum_{i \in I} (\bar{\lambda}_i - \lambda_i) g_i(\bar{x}) = - \sum_{i \in I} \lambda_i g_i(\bar{x}) \geq 0.$$

Thus, $L(\bar{x}, \lambda) \leq L(\bar{x}, \bar{\lambda})$ for all $\lambda \in \mathbb{R}_+^m$; that is, (SPC1) holds true.

Conversely, we now suppose that $\bar{x} \in C$ and there exists $\bar{\lambda} \in \mathbb{R}^m$ such that (NC), (SPC), and (CC) hold. We associate with $\bar{\lambda}$ the vectors $\lambda^i := \bar{\lambda} + e_i$, $i = 1, \dots, m$. By (NC), we have $\lambda^i \in \mathbb{R}_+^m$, $i \in I$. Further, by (SPC1),

$$0 \geq L(\bar{x}, \lambda^i) - L(\bar{x}, \bar{\lambda}) = g_i(\bar{x}), \quad \forall i \in I,$$

so $\bar{x} \in F$. From (SPC1) and (CC), we get

$$f(\bar{x}) = L(\bar{x}, 0_m) \leq L(\bar{x}, \bar{\lambda}) = f(\bar{x}) + \sum_{i \in I} \bar{\lambda}_i g_i(\bar{x}) = f(\bar{x}),$$

so we get $f(\bar{x}) = L(\bar{x}, \bar{\lambda})$. We thus have, by (SPC2),

$$\begin{aligned} f(\bar{x}) &= L(\bar{x}, \bar{\lambda}) = \inf \{L(x, \bar{\lambda}) : x \in C\} \\ &\leq \inf \{L(x, \bar{\lambda}) : x \in C, g_i(x) \leq 0, \forall i \in I\} \\ &= \inf \left\{ f(x) + \sum_{i \in I} \bar{\lambda}_i g_i(x) : x \in C, g_i(x) \leq 0, \forall i \in I \right\} \\ &\leq \inf \{f(x) : x \in C, g_i(x) \leq 0, \forall i \in I\} = v(P). \end{aligned}$$

Hence, $\bar{x} \in F^*$. The proof is complete. \square

What Theorem 4.43 asserts is that, under the assumptions on P (boundedness and SCQ), a feasible solution is optimal if and only if there exists a vector $\bar{\lambda} \in \mathbb{R}^m$ such that (NC), (SPC), and (CC) hold, in which case we say that $\bar{\lambda}$ is a *Lagrange vector*.

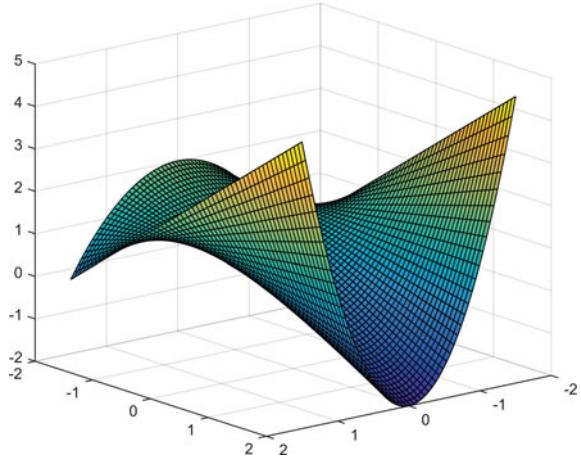
The next simple example, where $n = m = 1$, allows to visualize the saddle point of a convex optimization problem as $\text{gph } L \subset \mathbb{R}^3$.

Example 4.45 Consider

$$\begin{aligned} P : \text{Min } f(x) &= x^2 \\ \text{s.t. } x^2 - 1 &\leq 0, \\ x &\in \mathbb{R}. \end{aligned}$$

It is easy to see that $F^* = \{0_2\}$, $v(P) = 0$, $\mathcal{F}(z) = [-\sqrt{z+1}, \sqrt{z+1}]$ and $\vartheta(z) = 0$, if $z \geq -1$, while $\mathcal{F}(z) = \emptyset$ and $\vartheta(z) = +\infty$, if $z < -1$. Since

Fig. 4.21 $(0, 0)$ is a saddle point of L in Example 4.45



$\text{epi } \vartheta = [-1, +\infty[\times \mathbb{R}_+$, the unique sensitivity vector is 0. Figure 4.21 shows that $(0, 0)$ is a saddle point for $L(x, \lambda) = x^2 + \lambda(x^2 - 1)$ on $\mathbb{R} \times \mathbb{R}_+$ (observe that $L(x, \cdot)$ is an affine function on \mathbb{R}_+ for all $x \in \mathbb{R}$).

We now show that the condition $-\nabla f(\bar{x}) \in A(\bar{x}) = \text{cone}\{\nabla g_i(\bar{x}), i \in I(\bar{x})\}$ (the active cone at \bar{x} defined in (4.7)) also characterizes the optimality of $\bar{x} \in F$ when P is a differentiable convex optimization problem and SCQ holds (which means that the assumptions of the KKT Theorems 4.20 and 4.46 are independent of each other).

Theorem 4.46 (KKT theorem with convex constraints) *Assume that P is a bounded convex problem satisfying SCQ, with $f, g_i, i \in I$, differentiable on some open set containing C . Let $\bar{x} \in F \cap \text{int } C$. Then, the following statements are equivalent:*

(i) $\bar{x} \in F^*$.

(ii) $-\nabla f(\bar{x}) \in A(\bar{x})$.

(iii) There exists some $\bar{\lambda} \in \mathbb{R}^m$ such that:

$$(NC) \quad \bar{\lambda} \in \mathbb{R}_+^m;$$

$$(SC) \quad \nabla_x L(\bar{x}, \bar{\lambda}) = \nabla f(\bar{x}) + \sum_{i \in I} \bar{\lambda}_i \nabla g_i(\bar{x}) = 0_n; \text{ and}$$

$$(CC) \quad \bar{\lambda}_i g_i(\bar{x}) = 0, \quad \forall i \in I.$$

Proof Since (ii) and (iii) are trivially equivalent, it is sufficient to prove that (i) \Leftrightarrow (iii).

Assume that $\bar{x} \in F^*$. By Theorem 4.43, there exists a sensitivity vector $\bar{\lambda} \in \mathbb{R}_+^m$ satisfying (CC) and (SPC). As the second inequality in (SPC) is

$$L(\bar{x}, \bar{\lambda}) \leq L(x, \bar{\lambda}), \quad \forall x \in C,$$

we have that $\bar{x} \in \text{int } C$ is a global minimum of $x \mapsto L(x, \bar{\lambda})$ on C . Then, the Fermat principle implies that

$$\nabla_x L(\bar{x}, \bar{\lambda}) = \nabla f(\bar{x}) + \sum_{i \in I} \bar{\lambda}_i \nabla g_i(\bar{x}) = 0_n,$$

that is, (SC) holds.

We now assume the existence of $\bar{\lambda} \in \mathbb{R}^m$ such that (NC), (SC), and (CC) hold. Let $x \in F$. Due to the convexity of f and g_i , $i \in I$, we have

$$\begin{aligned} f(x) &\geq f(\bar{x}) + \sum_{i \in I} \bar{\lambda}_i g_i(x) \\ &\geq (f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x})) + \sum_{i \in I} \bar{\lambda}_i (g_i(\bar{x}) + \nabla g_i(\bar{x})^T (x - \bar{x})) \\ &= \left(f(\bar{x}) + \sum_{i \in I} \bar{\lambda}_i g_i(\bar{x}) \right) + \left(\nabla f(\bar{x}) + \sum_{i \in I} \bar{\lambda}_i \nabla g_i(\bar{x}) \right)^T (x - \bar{x}) \\ &= f(\bar{x}). \end{aligned}$$

Hence, $\bar{x} \in F^*$. \square

A vector $\bar{\lambda} \in \mathbb{R}^m$ such that (NC), (SC), and (CC) hold is said to be a *KKT vector*.

Remark 4.47 Observe that SCQ was not used in the proof of (iii) \Rightarrow (i). Therefore, in a convex problem, the existence of a KKT vector associated with \bar{x} implies the global optimality of \bar{x} .

Revisiting Example 4.37, $(\bar{x}^T, \bar{\lambda}) = (0, 0, 1)$ does not satisfy (SC) of Theorem 4.46 as f is not even differentiable at 0_2 . Concerning Example 4.45, it is easy to check that 0_2 satisfies (NC), (SC), and (CC).

The KKT conditions are used to either confirm or reject the optimality of a given feasible solution (e.g., the current iterate for some convex optimization algorithms) or as a filter allowing to elaborate a list of candidates to global minima. Under the assumptions of Theorem 4.46, the sensitivity theorem guarantees the existence of some sensitivity vector and the proofs of the saddle point and the KKT theorems show that such a vector is a KKT vector. So, if P has a unique KKT vector, then it is a sensitivity vector too.

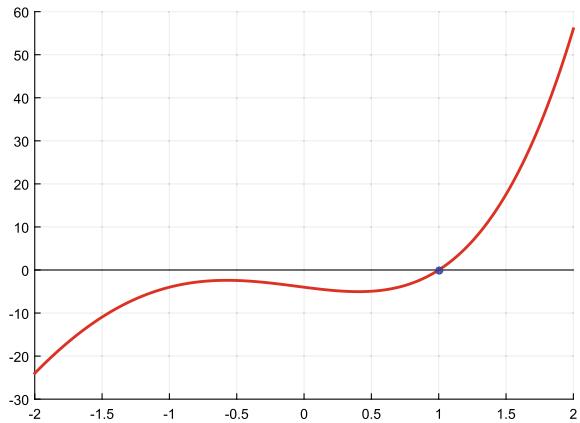
Example 4.48 We try to solve the optimization problem

$$\begin{aligned} P : \text{Min } f(x) &= 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2 \\ \text{s.t. } g_1(x) &= x_1^2 + x_2^2 - 5 \leq 0, \\ g_2(x) &= 3x_1 + x_2 - 6 \leq 0. \end{aligned}$$

We have $F \subset \sqrt{5}\mathbb{B}$, so F is compact,

$$\nabla f(x) = \begin{pmatrix} 4x_1 + 2x_2 - 10 \\ 2x_1 + 2x_2 - 10 \end{pmatrix}, \quad \nabla g_1(x) = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} \quad \text{and} \quad \nabla g_2(x) = \begin{pmatrix} 3 \\ 1 \end{pmatrix},$$

Fig. 4.22 Plot of the polynomial
 $\lambda_1^4 + 6\lambda_1^3 + \lambda_1^2 - 4\lambda_1 - 4$



with $\nabla^2 f$ and $\nabla^2 g_1$ positive definite while $\nabla^2 g_2$ is positive semidefinite. Thus, the three functions are convex and differentiable on \mathbb{R}^2 and P has a unique optimal solution. Moreover, 0_2 is a Slater point, so SCQ holds. There are four possible values for $I(x)$ (the parts of $I = \{1, 2\}$):

- $I(x) = \emptyset$: The unique solution to $\nabla f(x) = 0_2$ is $x = (0, 5)^T \notin F$.
- $I(x) = \{1\}$: We must solve the nonlinear system

$$\left\{ \begin{pmatrix} 4x_1 + 2x_2 - 10 \\ 2x_1 + 2x_2 - 10 \end{pmatrix} + \lambda_1 \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} = 0_2; x_1^2 + x_2^2 = 5 \right\}.$$

Eliminating x_1 and x_2 , one gets $\lambda_1^4 + 6\lambda_1^3 + \lambda_1^2 - 4\lambda_1 - 4 = 0$, whose unique positive root is $\lambda_1 = 1$ (see Fig. 4.22).

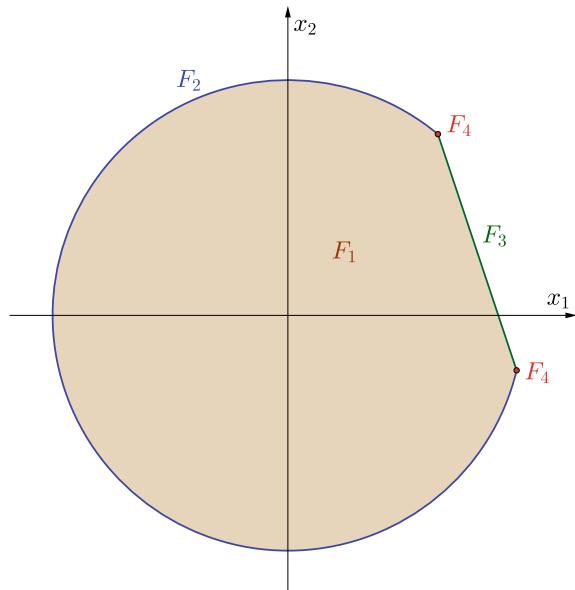
Replacing this value, we obtain $x^1 = (1, 2)^T \in F$, which is a minimum of P with corresponding KKT vector $(1, 0)^T$. Once we have obtained the unique minimum of P , there is no need to complete the discussion of the remaining cases $I(x) = \{2\}$ and $I(x) = \{1, 2\}$.

Figure 4.23 shows the partition of F into the sets F_1, F_2, F_3, F_4 corresponding to the parts of $I(x)$, $\emptyset, \{1\}, \{2\}$, and $\{1, 2\}$, respectively. Observe that $F_1 = \text{int } F$, F_2 is an arch of circle without its end points, F_3 is a segment without its extreme points, and, finally, F_4 is formed by two isolated points (the end points of the mentioned segment).

When P has multiple solutions, one can use the next result, which directly proves that the Lagrange vectors are sensitivity vectors.

Theorem 4.49 (Sensitivity and saddle points) *Let $\bar{x} \in F^*$. If $\bar{\lambda} \in \mathbb{R}^m$ satisfies (NC) $\bar{\lambda} \in \mathbb{R}_+^m$, (SPC) $L(\bar{x}, \lambda) \leq L(\bar{x}, \bar{\lambda}) \leq L(x, \bar{\lambda})$, $\forall x \in C, \lambda \in \mathbb{R}_+^m$, and (CC) $\bar{\lambda}_i g_i(\bar{x}) = 0$, $\forall i \in I$, then $\bar{\lambda}$ is a sensitivity vector for P , i.e.,*

Fig. 4.23 Applying the KKT conditions: partition of F depending on $I(x)$



$$\vartheta(z) \geq \vartheta(0_m) - \bar{\lambda}^T z, \quad \forall z \in \text{dom } \vartheta.$$

Proof Let $\bar{x} \in F^*$ and $\bar{\lambda} \in \mathbb{R}^m$ be such that (NC), (SPC), and (CC) hold. By (CC),

$$v(P) = f(\bar{x}) = f(\bar{x}) + \sum_{i \in I} \bar{\lambda}_i g_i(\bar{x}) = L(\bar{x}, \bar{\lambda}),$$

and, by (SPC),

$$v(P) = L(\bar{x}, \bar{\lambda}) \leq L(x, \bar{\lambda}) = f(x) + \sum_{i \in I} \bar{\lambda}_i g_i(x), \quad \forall x \in C.$$

Let $z \in \text{dom } \vartheta$ and $x \in C$ such that $g_i(x) \leq z, i \in I$. Then, by (NC),

$$v(P) \leq f(x) + \sum_{i \in I} \bar{\lambda}_i g_i(x) \leq f(x) + \bar{\lambda}^T z, \quad \forall x \in \mathcal{F}(z). \quad (4.29)$$

Finally, taking the infimum on $\mathcal{F}(z)$ of both members of (4.29), we get

$$\vartheta(0_m) = v(P) \leq \vartheta(z) + \bar{\lambda}^T z$$

This completes the proof. □

4.4.3 Lagrange Duality

We have seen in the last subsection that $h(y) := \inf_{x \in C} L(x, y)$, with $y \in \mathbb{R}_+^m$, provides a lower bound for $v(P)$. Thus, we associate with P the following problem, called *Lagrange dual* (or *Lagrangian dual*) of P ,

$$\begin{aligned} D^L : \text{Max } h(y) &:= \inf_{x \in C} L(x, y) \\ \text{s.t. } y &\in \mathbb{R}_+^m, \end{aligned}$$

whose optimal value is denoted by $v(D^L) \in \overline{\mathbb{R}}$. It is worth observing that $L(x, \cdot)$ is an affine function for all $x \in C$, so $h(\cdot) := \inf_{x \in C} L(x, \cdot)$ is a concave function and D^L is equivalent to a linearly constrained convex optimization problem. Hence, its feasible set $G = \mathbb{R}_+^m$ and its optimal set G^* are both convex subsets of \mathbb{R}^m . The weak duality theorem $v(D^L) \leq v(P)$ is a straightforward consequence of the definition of D^L (observe that the weak duality holds even in nonconvex optimization). The difference $v(P) - v(D^L) \geq 0$ is called the *duality gap* of the dual pair $P - D^L$.

The main result in any duality theory establishes that $v(D^L) = v(P)$ under suitable assumptions. This equation allows to certify the optimality of the current iterate or to stop the execution of primal-dual algorithms whenever an ε -optimal solution has been attained. *Strong duality* holds when, in addition to $v(D^L) = v(P)$, $G^* \neq \emptyset$. In linear optimization, it is known that the simultaneous consistency of the primal problem P and its dual one D^L guarantees that $v(D^L) = v(P)$ with $F^* \neq \emptyset$ and $G^* \neq \emptyset$. This is not the case in convex optimization, where strong duality requires the additional condition that SCQ holds (which is not enough in nonconvex optimization).

Theorem 4.50 (Strong Lagrange duality) *If P satisfies SCQ and it is bounded, then $v(D^L) = v(P)$ and $G^* \neq \emptyset$.*

Proof The assumptions guarantee, by Theorem 4.40, the existence of a sensitivity vector $\bar{y} \in \mathbb{R}_+^m$, and this vector satisfies $h(\bar{y}) = v(P)$ by Theorem 4.42. Then, $v(P) = h(\bar{y}) \leq v(D^L)$ and the conclusion follows from the weak duality theorem. \square

Therefore, in the simple Example 4.45, where SCQ holds, one has $h(y) = \inf_{x \in \mathbb{R}} (x^2 + y(x^2 - 1)) = -y$, for all $y \in \mathbb{R}_+^m$. So, $v(D^L) = \sup_{y \in \mathbb{R}_+} h(y) = 0 = v(P)$ and the optimal value of D^L is attained at $\bar{y} = 0$.

Let us revisit Example 4.37. There, the value of $h : \mathbb{R}_+ \rightarrow \overline{\mathbb{R}}$ at $y \in \mathbb{R}_+^m$ is

$$\begin{aligned} h(y) &= \inf_{x \in \mathbb{R} \times \mathbb{R}_+} (|x_1| + x_2 + yx_1) \\ &= \min \left\{ \inf_{x \in \mathbb{R}_+^2} ((y+1)x_1 + x_2), \inf_{x \in (-\mathbb{R}_+) \times \mathbb{R}_+} ((y-1) + x_2) \right\} \\ &= \begin{cases} 0, & 0 \leq y \leq 1 \\ -\infty, & y > 1 \end{cases}, \end{aligned}$$

as $\inf_{x \in \mathbb{R}_+^2} ((y+1)x_1 + x_2) = 0$. So, $v(D^L) = 0$ with optimal set $G^* = [0, 1]$.

We now obtain an explicit expression for the Lagrange dual of the quadratic problem with inequalities

$$\begin{aligned} P_Q : \text{Min } f(x) &= \frac{1}{2}x^T Qx - c^T x \\ \text{s.t. } a_i^T x &\leq b_i, \quad i \in I, \end{aligned}$$

considered in Subsection 4.3.2, assuming that Q is positive definite. Denote by A the $m \times n$ matrix whose rows are a_1^T, \dots, a_m^T and $b = (b_1, \dots, b_m)^T$. The Lagrange function of P_Q is

$$L_Q(x, y) = \frac{1}{2}x^T Qx + (-c + A^T y)^T x - b^T y.$$

Since $L_Q(\cdot, y)$ is strongly convex, its minimum on \mathbb{R}^n is attained at the unique zero of $\nabla_x L_Q(x, y) = Qx - c + A^T y$, i.e., at $Q^{-1}(c - A^T y)$. Thus,

$$\begin{aligned} h(y) &:= L_Q(Q^{-1}(c - A^T y), y) \\ &= -\frac{1}{2}y^T (AQ^{-1}A^T)y + (AQ^{-1}c - b)^T y - \frac{1}{2}c^T Q^{-1}c, \end{aligned}$$

which is a concave function. Hence, the Lagrange dual problem of P_Q ,

$$\begin{aligned} D_Q^L : \text{Max } h(y) \\ \text{s.t. } y_i \geq 0, \quad i = 1, \dots, m, \end{aligned}$$

is a convex quadratic problem as well, with very simple linear constraints (usually much simpler than those of P_Q). If SCQ holds, then $v(P_Q) = v(D_Q^L)$, with $G^* \neq \emptyset$, and one can exploit this simplicity in two different ways:

- Obtaining, as in linear optimization, an exact optimal solution of D_Q^L by means of some quadratic solver, and then the aimed optimal solution of P_Q by using (SC) and (CC);
- Interrupting the execution of any primal-dual algorithm whenever $f(x_k) - h(y_k) < \varepsilon$ for some tolerance $\varepsilon > 0$ (approximate stopping rule), as shown in Fig. 4.24.

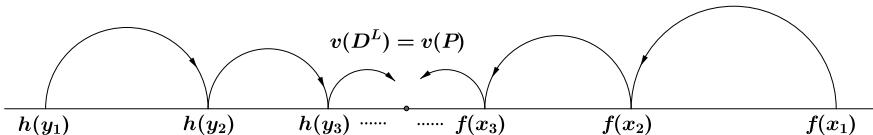


Fig. 4.24 Computing an ε -optimal solution

4.4.4 Wolfe Duality

Ph. Wolfe, an expert in quadratic optimization, proposed in 1961 [90] an alternative dual problem that allowed to cope with convex quadratic optimization problems whose objective function fails to be strongly convex.

Let P be an optimization problem of type (4.20), with convex and differentiable objective and constraint functions, and $C = \mathbb{R}^n$. The *Wolfe dual* of P is

$$\begin{aligned} D^W : & \text{Max } L(u, y) \\ \text{s.t. } & \nabla_u L(u, y) = 0_n, \\ & y_i \geq 0, \quad i = 1, \dots, m, \end{aligned}$$

where L denotes the Lagrange function of P . As for the previous dual problems, we denote by G and G^* the feasible and the optimal sets of D^W , respectively. In contrast to the pair $P - D^L$, the weak duality of $P - D^W$ is not immediate and follows from the convexity assumption.

Proposition 4.51 (Weak duality) *It holds that $v(D^W) \leq v(P)$.*

Proof Let $x \in F$ and $(u, y) \in G$. By the characterization of convex differentiable functions in Proposition 2.40,

$$f(x) - f(u) \geq \nabla f(u)^T(x - u) \quad (4.30)$$

and

$$g_i(x) - g_i(u) \geq \nabla g_i(u)^T(x - u), \quad i = 1, \dots, m. \quad (4.31)$$

From (4.30), the definition of G , inequality (4.31), and the primal feasibility of x , in this order, one gets

$$\begin{aligned} f(x) - f(u) &\geq \nabla f(u)^T(x - u) \\ &= -\sum_{i=1}^m y_i \nabla g_i(u)^T(x - u) \\ &\geq -\sum_{i=1}^m y_i(g_i(x) - g_i(u)) \\ &\geq \sum_{i=1}^m y_i g_i(u), \end{aligned}$$

so that

$$f(x) \geq f(u) + \sum_{i=1}^m y_i g_i(u) = L(u, y),$$

showing that $v(P) \geq v(D^W)$. □

We cannot associate a Wolfe dual to the problem of Example 4.37 as f is not differentiable. Regarding the problem P in Example 4.45, $L(u, y) = (1+y)u^2 - y$ and

$$G = \{(u, y) \in \mathbb{R} \times \mathbb{R}_+ : (1+y)u = 0\} = \{0\} \times \mathbb{R}_+.$$

So, $G^* = \{0_2\}$ and $v(D^W) = v(P)$. We now prove that, in fact, the fulfillment of the strong duality property in this example is a consequence of SCQ.

Theorem 4.52 (Strong Wolfe duality) *If P is solvable and SCQ holds, then $v(D^W) = v(P)$ and D^W is solvable.*

Proof Let $\bar{x} \in F^*$. It will be sufficient to show the existence of some $\bar{y} \in \mathbb{R}_+^m$ such that $(\bar{x}, \bar{y}) \in G^*$ and $f(\bar{x}) = L(\bar{x}, \bar{y})$.

We first observe that, by the convexity assumption and Proposition 2.40,

$$L(u, y) - L(x, y) \geq \nabla_u L(x, y)^T (u - x), \quad \forall y \in \mathbb{R}_+^m, \forall u, x \in \mathbb{R}^n. \quad (4.32)$$

Let $(u_1, y), (u_2, y) \in G$. Applying (4.32) to the couples (u_1, y) and (u_2, y) , since $\nabla_u L(u_j, y) = 0_n$, $j = 1, 2$, one gets $L(u_1, y) - L(u_2, y) \geq 0$ and $L(u_2, y) - L(u_1, y) \geq 0$; i.e., $L(u_1, y) = L(u_2, y)$. Thus, $L(\cdot, y)$ is constant on G for all $y \in \mathbb{R}_+^m$.

By the saddle point Theorem 4.43, with $C = \mathbb{R}^n$, there exists $\bar{y} \in \mathbb{R}_+^m$ such that

$$L(\bar{x}, y) \leq L(\bar{x}, \bar{y}) \leq L(x, \bar{y}), \quad \forall x \in \mathbb{R}^n, y \in \mathbb{R}_+^m, \quad (4.33)$$

and

$$\bar{y}_i g_i(\bar{x}) = 0, \quad i = 1, \dots, m. \quad (4.34)$$

Thus, \bar{x} is a global minimum of $x \mapsto L(x, \bar{y})$, so the Fermat principle implies that $\nabla_x L(\bar{x}, \bar{y}) = 0_n$; i.e., $(\bar{x}, \bar{y}) \in G$.

From (4.33) and the constancy of $L(\cdot, y)$ on G , we have

$$\begin{aligned} L(\bar{x}, \bar{y}) &= \max\{L(\bar{x}, y) : y \in \mathbb{R}_+^m\} \\ &\geq \max\{L(\bar{x}, y) : (\bar{x}, y) \in G\} \\ &= \max\{L(x, y) : (x, y) \in G\} \\ &\geq L(\bar{x}, \bar{y}), \end{aligned}$$

so that $(\bar{x}, \bar{y}) \in G^*$, with

$$L(\bar{x}, \bar{y}) = f(\bar{x}) + \sum_{i=1}^m \bar{y}_i g_i(\bar{x}) = f(\bar{x})$$

thanks to (4.34). \square

Generally speaking, D^W may be difficult to deal with computationally, because its objective function is not concave in both variables. However, in convex quadratic optimization, D^W is preferable to D^L when Q is only positive semidefinite. Indeed, let

$$\begin{aligned} P_Q : \text{Min } f(x) &= \frac{1}{2}x^T Qx - c^T x \\ \text{s.t. } a_i^T x &\leq b_i, \quad i \in I, \end{aligned}$$

be a convex quadratic problem, and let A be the $m \times n$ matrix whose rows are a_1^T, \dots, a_m^T . Since its Lagrange function is

$$L_Q(u, y) = \frac{1}{2}u^T Qu + (-c + A^T y)^T u - b^T y,$$

the u -gradient reads

$$\nabla_u L_Q(u, y) = Qu + (-c + A^T y).$$

Then,

$$G = \{(u, y) \in \mathbb{R}^n \times \mathbb{R}_+^m : -c + A^T y = -Qu\}$$

and the Wolfe dual of P_Q can be expressed as

$$\begin{aligned} D_Q^W : \text{Max } & -\frac{1}{2}u^T Qu - b^T y \\ \text{s.t. } & -c + A^T y = -Qu \\ & y_i \geq 0, \quad i = 1, \dots, m, \end{aligned}$$

whose objective function is concave (but not strongly concave) and its linear constraints are very simple, while D_Q^L does not have an explicit expression as Q is not invertible.

4.5 A Glimpse on Conic Optimization*

We have already found in Subsection 4.3.2 convex feasible sets of the form $F = \{x \in \mathbb{R}^n : Ax + b \in K\}$, where A is an $m \times n$ matrix, $b \in \mathbb{R}^m$ and K is a polyhedral convex cone in \mathbb{R}^m . The following three convex cones frequently arise in practice:

- The *positive orthant* of \mathbb{R}^m , \mathbb{R}_+^m , which converts the conic constraint $Ax + b \in K$ into an ordinary linear inequality system;
- The *second-order cone*, also called the ice-cream cone,

$$K_p^m := \{x \in \mathbb{R}^m : x_m \geq \|(x_1, \dots, x_{m-1})^T\|\},$$

and cartesian products of the form $\prod_{j=1}^l K_p^{m_j+1}$ (K_p^3 is represented twice in Fig. 4.6).

- The *cone of positive semidefinite symmetric matrices* in \mathcal{S}_q , usually denoted by \mathcal{S}_q^+ . Here, we identify the space of all $q \times q$ symmetric matrices \mathcal{S}_q with $\mathbb{R}^{q(q+1)/2} = \mathbb{R}^m$. From now on, we write $A \succeq 0$ when $A \in \mathcal{S}_q$ is positive semidefinite and $A \succ 0$ when it is positive definite.

It is easy to see that the above cones are closed and convex, and they have nonempty interior (observe that $A \succ 0$ implies that $A \in \text{int } S_q^+$ as the eigenvalues of A are continuous functions of its entries). Moreover, they are *pointed* and *symmetric* (meaning that $K \cap -K = \{0_m\}$ and that $K^\circ = -K$, respectively). The pointedness of \mathbb{R}_+^m and K_p^m is evident, while, for S_q^+ , it follows from the characterization of the symmetric positive semidefinite matrices by the nonnegativity of their eigenvalues. The symmetry is also evident for \mathbb{R}_+^m , it can be easily proved from the Cauchy–Schwarz inequality for K_p^m , and it is a nontrivial result proved by Moutard for S_q^+ [58, Theorem 7.5.4].

We first consider a *linear conic problem* of the form

$$\begin{aligned} P_K : \text{Min } & c^T x \\ \text{s.t. } & Ax + b \in K, \end{aligned}$$

where $c \in \mathbb{R}^n$, A is an $m \times n$ matrix, $b \in \mathbb{R}^m$, and K is a pointed closed convex cone in \mathbb{R}^m such that $\text{int } K \neq \emptyset$. The assumptions on K guarantee that K° satisfies the same properties and the existence of a *compact base* of K° , i.e., a compact convex set W such that $0_m \notin W$ and $K^\circ = \text{cone } W$ [40, page p. 447]. For instance, compact bases of K° are $W = -\text{conv}\{e_1, \dots, e_m\}$, for $K = \mathbb{R}_+^m$, and $W = \mathbb{B} \times \{-1\}$, for $K = K_p^m$. By the same argument as in Subsection 4.3.2, since $K^\circ = \text{cone } W$,

$$Ax + b \in K \iff w^T(Ax + b) \leq 0, \quad \forall w \in W. \quad (4.35)$$

So, the conic constraint $Ax + b \in K$ can be replaced in P_K by the right-hand side of (4.35), getting the equivalent *linear semi-infinite problem*

$$\begin{aligned} P_K^1 : \text{Min } & c^T x \\ \text{s.t. } & (w^T A)x \leq -w^T b, \quad w \in W, \end{aligned}$$

(a linear optimization problem with infinitely many inequalities), whose theory, methods, and applications are exposed in [43] (see also [44]). From the existence theorem [43, Corollary 3.1.1] for this type of problems,

$$F \neq \emptyset \iff \begin{pmatrix} 0_n \\ -1 \end{pmatrix} \notin \text{cone}\{[A \mid -b]^T w, w \in W\},$$

in which case, optimal solutions can be computed by the corresponding numerical methods. Stopping rules for primal-dual algorithms can be obtained from the *Haar dual problem* of P_K^1 ,

$$\begin{aligned} D_K^1 : \text{Max } & \sum_{w \in W} b^T w \lambda_w \\ \text{s.t. } & \sum_{w \in W} \lambda_w A^T w = -c, \\ & \lambda \in \mathbb{R}_+^{(W)}, \end{aligned}$$

where $\mathbb{R}_+^{(W)}$ denotes the convex cone of real-valued functions on W which vanish except on a finite subset of W . One always has the weak inequality $v(D_K^1) \leq v(P_K^1)$ as, given a pair (x, λ) of primal-dual feasible solutions,

$$c^T x = - \sum_{w \in W} \lambda_w (A^T w)^T x \geq \sum_{w \in W} w^T b \lambda_w.$$

Moreover, the strong duality holds under the *Slater constraint qualification* (still abbreviated as SCQ) that there exists some $\hat{x} \in \mathbb{R}^n$ such that $(w^T A)\hat{x} < -w^T b$ for all $w \in W$ [44]. Taking into account the compactness of the index set in P_K^1 and the continuity of the coefficients with respect to the index, it is also possible to associate with P_K^1 another dual problem, say D_K^2 , which is also linear and is posed on a certain infinite dimensional space of measures on W that is called *continuous dual* [4]. Once again, strong duality holds for D_K^2 under the same SCQ as above.

We can also reformulate P_K^1 as a convex optimization problem with a single constraint function $g(x) := \max_{w \in W} \{(w^T A)x + w^T b\}$ for all $x \in \mathbb{R}^n$. Obviously, P_K^1 is equivalent to

$$\begin{aligned} P_K^2 : \text{Min } f(x) &= c^T x \\ \text{s.t. } g(x) &\leq 0, \end{aligned}$$

whose Lagrange dual is

$$\begin{aligned} D_K^2 : \text{Max } h(y) &:= \inf_{x \in \mathbb{R}^n} \{c^T x + yg(x)\} \\ \text{s.t. } y &\geq 0. \end{aligned}$$

Due to the lack of differentiability of g , the Wolfe dual of P_K^2 is not well defined.

Observe that the dual pair $P_K^1 - D_K^1$ is preferable to the pair $P_K^2 - D_K^2$ because:

- The Slater constraint qualification for the pair $P_K^1 - D_K^1$ is weaker than the one corresponding to the pair $P_K^2 - D_K^2$, i.e., the existence of some $\hat{x} \in \mathbb{R}^n$ such that $g(\hat{x}) < 0$.
- D_K^2 can hardly be solved in practice as no explicit expression of g is available except for particular cases.

However, the preferable dual of P_K , from a computational point of view, is the so-called *conic dual problem*:

$$\begin{aligned} D_K : \text{Max } b^T y \\ \text{s.t. } A^T y &= -c, \\ y &\in K^\circ. \end{aligned}$$

To check the weak duality, take an arbitrary primal-dual feasible solution (x, y) . Then, since $Ax + b \in K$, we have

$$c^T x = -(A^T y)^T x = -y^T Ax = -y^T(Ax + b) + b^T y \geq b^T y.$$

The strong duality holds for the pair $P_K - D_K$ when there exists some $\hat{x} \in \mathbb{R}^n$ such that $A\hat{x} + b$ belongs to the *relative interior* of K , i.e., to the intersection of K with some open set of \mathbb{R}^n , which is the peculiar form of SCQ for this class of linear conic problems [74].

In order to motivate the definition of a more general class of conic problems, let us represent by x_1 , instead of x , the decision variable in P_K and introduce a new variable $x_2 := Ax_1 + b$. Denoting $\tilde{c} := \begin{pmatrix} c \\ 0_m \end{pmatrix}$ and $x := \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^{n+m}$, problem P_K turns out to be equivalent to

$$\begin{aligned} P_K^3 : \text{Min } & f(x) = \tilde{c}^T x \\ \text{s.t. } & [A \mid -I_n]x = -b \\ & x \in \mathbb{R}^n \times K. \end{aligned}$$

Obviously, P_K^3 is a particular case of the so-called *general linear conic problem* [6]

$$\begin{aligned} P_{\mathcal{K}} : \text{Min } & \langle c, x \rangle \\ \text{s.t. } & \langle a_i, x \rangle = b_i, \quad i \in I = \{1, \dots, m\}, \\ & x \in \mathcal{K}, \end{aligned}$$

where \mathcal{K} is a convex cone in a given linear space \mathcal{Z} equipped with an inner product $\langle \cdot, \cdot \rangle$, $c, a_1, \dots, a_m \in \mathcal{Z}$, and $b_1, \dots, b_m \in \mathbb{R}$ (observe that the definitions of convex set, convex function, cone, polar cone, and convex optimization problem make sense when the usual decision space \mathbb{R}^n is replaced by \mathcal{Z}). We associate with $P_{\mathcal{K}}$ the following linear conic problem:

$$\begin{aligned} D_{\mathcal{K}} : \text{Max } & b^T y \\ \text{s.t. } & \sum_{i \in I} y_i a_i + z = c, \\ & (y, z) \in \mathbb{R}^m \times \mathcal{K}^\circ, \end{aligned}$$

where $b = (b_1, \dots, b_m)^T$ and \mathcal{K}° is the polar cone of \mathcal{K} . It is easy to check the convexity of $P_{\mathcal{K}}$ and $D_{\mathcal{K}}$. Taking an arbitrary primal-dual feasible solution $(x, (y, z))$, one has

$$b^T y = \left\langle \sum_{i \in I} y_i a_i, x \right\rangle = \langle c - z, x \rangle = \langle c, x \rangle - \langle z, x \rangle \geq \langle c, x \rangle,$$

so that weak duality holds. For this reason, $D_{\mathcal{K}}$ is said to be the *conic dual problem* of $P_{\mathcal{K}}$. The way of guaranteeing strong duality for the pair $P_{\mathcal{K}} - D_{\mathcal{K}}$ depends on \mathcal{Z} and \mathcal{K} :

- In *linear optimization*, $\mathcal{Z} = \mathbb{R}^m$, $\langle c, x \rangle = c^T x$, and $\mathcal{K} = \mathbb{R}_+^m$. Strong duality holds just assuming the existence of some primal-dual feasible solution; i.e., no CQ is needed.
- In *second-order cone optimization*, $\mathcal{Z} = \prod_{j=1}^l \mathbb{R}^{n_j+1}$, $\langle c, x \rangle = \sum_{j=1}^l c_j^T x_j$, and $\mathcal{K} = \prod_{j=1}^l K_p^{n_j+1}$.
- In *semidefinite optimization*, $\mathcal{Z} = \mathcal{S}_n$, $\langle C, X \rangle$ is the trace of the product matrix CX , and $\mathcal{K} = \mathcal{S}_n^+$.

The SCQ in second-order cone optimization and semidefinite optimization reads as follows: There exists a primal-dual feasible solution $(\hat{x}, (\hat{y}, \hat{z}))$ such that $\hat{x}, \hat{z} \in \text{int } \mathcal{K}$. For the mentioned last two classes of optimization problems, SCQ also guarantees the existence of a primal optimal solution [6]. A favorable consequence of the symmetry of the three mentioned cones is that the corresponding conic problems admit efficient primal-dual algorithms [69].

It is worth mentioning that second-order cone optimization problems can be reformulated as semidefinite optimization ones as a consequence of the identity

$$K_p^m = \left\{ x \in \mathbb{R}^m : \begin{bmatrix} x_m & 0 & \dots & 0 & x_1 \\ 0 & x_m & & 0 & x_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & x_m & x_{m-1} \\ x_1 & x_2 & \dots & x_{m-1} & x_m \end{bmatrix} \succeq 0 \right\},$$

but these reformulations are not used in practice as the specific second-order cone optimization methods are much more efficient than the adaptations of the semidefinite optimization ones [3].

Many textbooks on convex optimization, e.g., [10, 15], pay particular attention to the theory and methods of conic optimization. Several chapters of [7] also deal with the theory and methods of conic and semidefinite optimization, while [2] is focused on second-order cone optimization. Regarding applications, [10, 15] present interesting applications to engineering and finance (e.g., the portfolio problem), while [85] contains chapters reviewing applications of conic optimization to nonlinear optimal control (pp. 121–133), truss topology design (pp. 135–147), and financial engineering (pp. 149–160).

4.6 Exercises

4.1 Prove that $\bar{x} = (1, 1)^T$ is the unique optimal solution to the problem

$$\begin{aligned} P : \text{Min } f(x) &= (x_1 - 2)^2 + (x_2 - 1)^2 \\ \text{s.t. } g_1(x) &= x_1^2 - x_2 \leq 0, \\ g_2(x) &= x_1^2 + x_2^2 - 2 \leq 0. \end{aligned}$$

4.2 Express a positive number a as the sum of three numbers so that the sum of its corresponding cubes is minimized, under the following assumptions:

- (a) The three numbers are arbitrary.
- (b) The three numbers are positive.
- (c) The three numbers are nonnegative.

4.3 Solve the following problem by using the KKT conditions

$$\begin{aligned} P : \text{Min } & e^{x^2+y^2} \\ \text{s.t. } & 2x - y = 4. \end{aligned}$$

4.4 Solve

$$\begin{aligned} P : \text{Min } & f(x_1, x_2) = x_1^{-1} x_2^{-1} \\ \text{s.t. } & x_1 + x_2 \leq 2, \\ & x_1 > 0, x_2 > 0. \end{aligned}$$

4.5 Solve the optimization problem posed in Exercise 1.3 by using the KKT conditions.

4.6 Determine whether it is true or false that $\bar{x} = \left(-\frac{1}{3}, -\frac{1}{3}, -\frac{1}{3}, -3\right)^T$ is, from all the solutions of the linear system

$$\left\{ \begin{array}{l} x_1 + x_2 - x_3 \geq -1 \\ 3x_1 + 3x_2 + 3x_3 - 4x_4 \geq 9 \\ -x_1 - x_2 - x_3 \geq 1 \\ -3x_1 + x_2 + x_3 \geq -1 \\ x_1 - 3x_2 + x_3 \geq -1 \\ -x_1 - x_2 + 3x_3 \geq -3 \end{array} \right\},$$

the closest one to the origin in \mathbb{R}^4 .

4.7 Consider the following problem in \mathbb{R}^2 :

$$\begin{aligned} P : \text{Min } & f(x) = 4x_1^2 + x_2^2 - 8x_1 - 4x_2 \\ \text{s.t. } & x_1 + x_2 \leq 4 \\ & 2x_1 + x_2 \leq 5 \\ & -x_1 + 4x_2 \geq 2 \\ & x_1 \geq 0, x_2 \geq 0. \end{aligned}$$

- (a) Reformulate P in such a way that the level curves are circumferences.
- (b) Solve P graphically.
- (c) Prove analytically that the result obtained in (b) is really true.
- (d) Solve the problem obtained when we add to P the constraint $x_2 \geq 3$.

4.8 Solve

$$\begin{aligned} P : \text{Max } & f(x) = 20x_1 + 16x_2 - 2x_1^2 - x_2^2 - x_3^2 \\ \text{s.t. } & x_1 + x_2 \leq 5, \\ & x_1 + x_2 - x_3 = 0, \\ & x_1 \geq 0, x_2 \geq 0. \end{aligned}$$

4.9 Consider the problem

$$\begin{aligned} P : \text{Min } & x_1^2 + x_2^2 + 5x_1 \\ \text{s.t. } & x_1^2 + x_2^2 \leq 4, \\ & -x_2 \leq -2. \end{aligned}$$

(a) Solve it graphically.

(b) Analyze the fulfillment of the KKT conditions at the point obtained in (a).

4.10 Check that the convex optimization problem

$$\begin{aligned} P : \text{Min } & f(x, y) = x \\ \text{s.t. } & x^2 + (y - 1)^2 \leq 1, \\ & x^2 + (y + 1)^2 \leq 1, \end{aligned}$$

has a unique global minimum that does not satisfy the KKT conditions. Why does such a thing happen?

4.11 Prove that the bounded problem

$$\begin{aligned} P : \text{Min } & f(x) = x_1^2 - 2x_1 + x_2^2 \\ \text{s.t. } & x_1 + x_2 \leq 0, \\ & x_1^2 - 4 \leq 0, \end{aligned}$$

has a unique optimal solution with a unique KKT vector associated with it. From that information, what can you say about the effect on the optimal value of small perturbations on the right-hand side of each constraint?

4.12 Consider the parametric problem

$$\begin{aligned} P(z) : \text{Min } & f(x) = x_1^2 - 2x_1 + x_2^2 + 4x_2 \\ \text{s.t. } & 2x_1 + x_2 \leq z, \\ & x \in \mathbb{R}^2. \end{aligned}$$

(a) Compute the value function ϑ .

(b) Check that ϑ is convex and differentiable at 0.

(c) Find the set of sensitivity vectors for $P(0)$.

4.13 Consider the problem

$$\begin{aligned} P : \text{Min } & f(x) = x_1^2 - x_1 x_2 \\ \text{s.t. } & x_1 - x_2 \leq 2, \end{aligned}$$

and its corresponding parametric problem

$$\begin{aligned} P(z) : \text{Min } & f(x) = x_1^2 - x_1 x_2 \\ \text{s.t. } & x_1 - x_2 - 2 \leq z. \end{aligned}$$

- (a) Identify the feasible solutions to P satisfying the KKT conditions, and determine whether any of them is an optimal solution.
 (b) Determine whether the value function is convex and whether P satisfies SCQ.
 (c) Compute a sensitivity vector.

4.14 The utility function of a consumer is $u(x, y) = xy$, where x and y denote the consumed quantities of two goods A and B, whose unit prices are 2 and 3 c.u., respectively. Maximize the consumer utility, knowing that she has 90 c.u.

4.15 A tetrahedron (or triangular pyramid) is *rectangular* when three of its faces are rectangle triangles that we will name *cateti*, whereas the fourth face will be named *hypotenuse*. Design a rectangular tetrahedron whose hypotenuse has minimum area being the pyramid height on the hypotenuse h meters.

4.16 Consider the convex optimization problem

$$\begin{aligned} P : \text{Min } f(x) &= e^{-x_2} \\ \text{s.t. } \|x\| - x_1 &\leq 0, \\ x &\in \mathbb{R}^2. \end{aligned}$$

- (a) Identify the feasible set multifunction \mathcal{F} and the value function ϑ .
 (b) Study the continuity and differentiability of ϑ on its domain.
 (c) Compute the sensitivity vectors.
 (d) Determine whether the optimal values of P and of its Lagrange dual problem D^L are equal.

4.17 Consider the convex optimization problem

$$\begin{aligned} P : \text{Min } f(x) &= \|x\| \\ \text{s.t. } x_1 + x_2 &\leq 0, \\ x &\in \mathbb{R}^2. \end{aligned}$$

- (a) Identify the feasible set multifunction \mathcal{F} and the value function ϑ .
 (b) Study the continuity and differentiability of ϑ on its domain.
 (c) Compute the sensitivity vectors of ϑ .
 (d) Compute the optimal set of P .
 (e) Compute the optimal set of its Lagrange dual problem D^L .
 (f) Determine whether strong duality holds.

4.18 Solve analytically and geometrically the problem

$$\begin{aligned} P : \text{Min } f(x) &= x_1^2 + x_2^2 + x_3^2 \\ \text{s.t. } x_1 + x_2 + x_3 &\leq -3, \end{aligned}$$

and compute its sensitivity vectors. Solve also the Lagrange dual D^L and the Wolfe dual D^W of P , showing that the strong duality holds for both duality pairs without using the SCQ.

4.19 Consider the problem

$$\begin{aligned} P : \text{Min } & f(x) = x \\ \text{s.t. } & g(x) = x^2 \leq 0. \end{aligned}$$

- (a) Solve P analytically, if possible.
- (b) Express analytically and represent graphically $\text{gph } \mathcal{F}$.
- (c) Identify the value function ϑ and represent graphically $\text{gph } \vartheta$.
- (d) Analyze the differentiability of ϑ on the interior of its domain.
- (e) Compute the sensitivity vectors of ϑ (here scalars, since $m = 1$).
- (f) Compute the saddle points of the Lagrange function L .
- (g) Compute the KKT vectors on the optimal solutions of P .
- (h) Check the strong duality property for the Lagrange dual D^L and for the Wolfe dual D^W of P .

Part II

Numerical Optimization

Chapter 5

Unconstrained Optimization Algorithms



In this chapter we will present the most representative algorithms for solving an optimization problem without functional constraints, that is, the problem

$$\begin{aligned} P : \text{Min } f(x) \\ \text{s.t. } x \in C, \end{aligned}$$

where $\emptyset \neq C \subset \text{dom } f \subset \mathbb{R}^n$. We will usually assume that f is smooth on C , i.e., that $f \in \mathcal{C}^1(V)$, where V is an open set such that $C \subset V \subset \text{dom } f$. In fact, in many cases, to simplify, we will assume that $C = \mathbb{R}^n$, which means that P is the problem (1.1) with neither functional nor constraint sets. We will present conceptual algorithms that generate infinite sequences, which can be converted into implementable algorithms by adding some stopping criteria as the ones inspired in errors introduced in Subsection 5.2.1 or some approximate optimality conditions.

Given an initial point x_0 , an algorithm must generate a sequence of points x_1, x_2, \dots . In order to decide how to derive the following iteration to x_k , the algorithms use information about f at x_k (and maybe also about the previous iterations x_0, \dots, x_{k-1}). Usually, this information cannot be obtained for free, and this is the reason why efficient algorithms with respect to the use of this information are preferred.

5.1 Line Search Methods

These type of algorithms are defined as follows. Given a point x_k , a direction p_k is chosen. The new point x_{k+1} is found by moving a step of size α_k in the direction p_k . That is, the next iteration is given by

$$x_{k+1} = x_k + \alpha_k p_k,$$

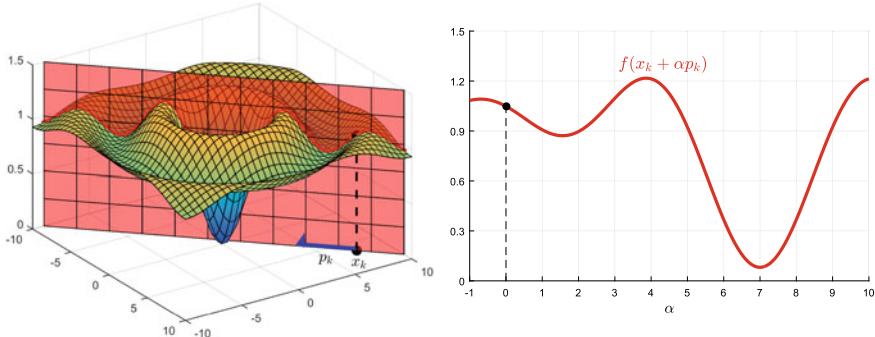


Fig. 5.1 Strategy of a line search method

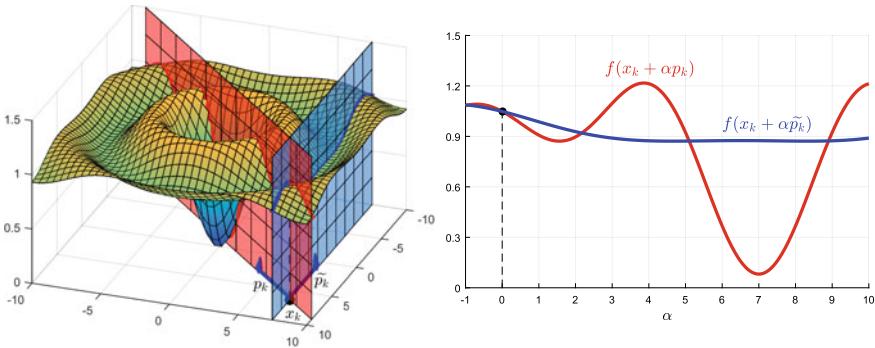


Fig. 5.2 The efficiency of a line search method will depend on both the direction and the stepsize chosen

where p_k is the *search direction* and α_k is the *stepsize*; see Fig. 5.1. Of course, the efficiency of the method will depend on both choices, as one can see in Fig. 5.2.

The stepsize α_k could be determined by solving the problem

$$\text{Min}_{\alpha > 0} f(x_k + \alpha p_k). \quad (5.1)$$

The one-dimensional optimization problem (5.1) is a global optimization problem, which is known as *exact line search*. With the exception of very few particular cases, the algorithms normally propose “reasonable” values of α_k (which approximately fulfill the objective in (5.1)). This strategy is referred as *inexact line search*.

Most of the line search methods use directions p_k that are of descent, a notion that we define next.

Definition 5.1 One says that p_k is a *descent direction* for the function f at x_k if

$$f'(x_k; p_k) = \nabla f(x_k)^T p_k < 0. \quad (5.2)$$

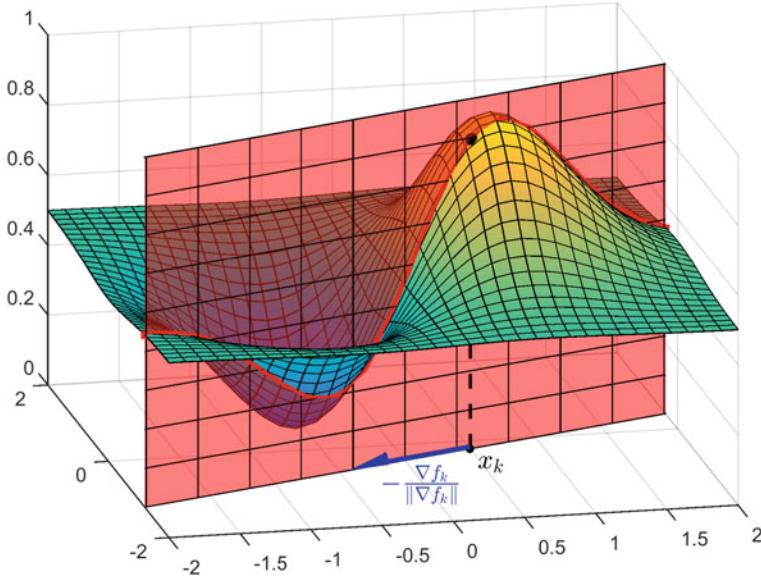


Fig. 5.3 Search direction of the steepest descent method

The directions p_k and \tilde{p}_k in Fig. 5.2 are both descent directions. Descent directions guarantee a decrease in the value of f when moving from x_k in the direction p_k : for $\alpha > 0$ sufficiently small, one has

$$f(x_{k+1}) = f(x_k + \alpha p_k) = f(x_k) + \alpha \nabla f(x_k)^T p_k + o(\alpha) < f(x_k).$$

To simplify, we will denote $\nabla f(x_k) \equiv \nabla f_k$.

The unitary direction of fastest decrease will be the solution to the problem

$$\text{Min}_{\|p\|=1} p^T \nabla f_k. \quad (5.3)$$

Since $p^T \nabla f_k = \|p\| \|\nabla f_k\| \cos \theta$, where θ is the angle between p and ∇f_k , we have that the minimum in (5.3) is attained when $\cos \theta$ has its minimum value -1 at $\theta = 180^\circ \equiv \pi$, that is, when

$$p = -\frac{\nabla f_k}{\|\nabla f_k\|}. \quad (5.4)$$

This is the (unitary) direction used by the so-called *steepest descent method*; see Fig. 5.3. As a consequence of (5.2), any direction forming an angle smaller than $90^\circ \equiv \pi/2$ with $-\nabla f_k$ will be a descent direction.

5.1.1 The Family of Gradient Methods

This name is given to the methods that use search directions of the form

$$p_k = -B_k^{-1} \nabla f_k,$$

where B_k is a nonsingular symmetric matrix. Observe that when B_k is positive definite, the above direction defines a descent direction: Indeed, if $\nabla f_k \neq 0$, one has

$$f'(x_k; p_k) = \nabla f_k^T p_k = -\nabla f_k^T B_k^{-1} \nabla f_k < 0.$$

The most important algorithms within this family are:

- $B_k = I_n$, in the steepest descent method (5.4);
- $B_k = \nabla^2 f_k$, in Newton's method;
- $B_k \approx \nabla^2 f_k$, in the so-called quasi-Newton methods.

Here $\nabla^2 f_k$ represents $\nabla^2 f(x_k)$.

The idea behind *Newton's method* is to minimize in each iteration the second-order approximation of $f(x_k + p)$:

$$f(x_k + p) \approx f_k + \nabla f_k^T p + \frac{1}{2} p^T \nabla^2 f_k p =: m_k(p).$$

If $\nabla^2 f_k$ is positive definite, then $m_k(p)$ is a coercive quadratic function (see Proposition 3.1), and the vector p obtained by minimizing $m_k(p)$ is the *Newton's direction*. Indeed, by making zero the derivative of $m_k(p)$, we find its explicit form:

$$p_k = -(\nabla^2 f_k)^{-1} \nabla f_k. \quad (5.5)$$

In the *pure Newton's method*, the stepsize is constantly chosen as $\alpha_k = 1$. Observe that this method finds the minimum in only one step when f is a quadratic function determined by a positive definite matrix. Most of the implementations of Newton's method choose a stepsize $\alpha = 1$ whenever possible, and only adjust its size if a satisfactory reduction in the value of f is not attained. When $\nabla^2 f_k$ is not positive definite, Newton's direction (5.5) may not exist or may not be a descent direction. If this is the case, there are different variations of Newton's method that would modify the search direction p_k to transform it into a descent direction.

5.1.2 The Stepsize

When computing the stepsize α_k in the descent direction p_k , we aim to balance two objectives. On the one hand, we would like to choose α_k in such a way that f is

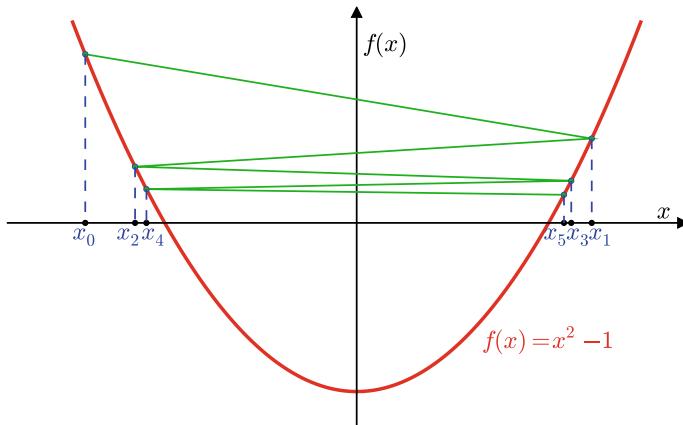


Fig. 5.4 $f(x_k) = \frac{1}{k}$ converges to 0 and not to the minimum value -1

substantially reduced, but on the other hand, we do not want to spend too much time on this. We know that the ideal choice would be a linear exact search, that is, a global minimizer of the univariate function $\phi(\cdot)$ defined by

$$\phi(\alpha) := f(x_k + \alpha p_k), \quad \alpha > 0. \quad (5.6)$$

Nevertheless, in general, the computational cost of such a search is unacceptable. Even finding a local minimum of ϕ with a moderate precision usually requires too many evaluations of f , and possibly of its gradient ∇f .

Inexact line searches aim to obtain an α_k allowing to reach an adequate reduction of f with a reasonable computational cost. These types of algorithms try a series of candidates for α_k , accepting one of these values when certain conditions are satisfied. The line search is performed in two phases: In the first one, an *interval* containing the desired stepsizes is obtained, and in a second phase of bisection or interpolation, a *good* stepsize in this interval is chosen. We analyze next this first phase of some inexact line search algorithms, and we show that effective stepsizes do not need to be close to the minima of the function $\phi(\alpha)$.

A simple condition that one can impose to α_k is that it should give us a reduction in f , i.e., that $f(x_k + \alpha_k p_k) < f(x_k)$. Nevertheless, as we can see in Fig. 5.4, this requirement is not sufficient: The (global) minimum of $f(x) = x^2 - 1$ is $f^* = -1$, and the sequence of values of the function $f(x_k) = 1/k$, for $k = 1, 2, \dots$, is strictly decreasing but converges to zero, not to the minimum value -1 .

The problem with the method applied in Fig. 5.4 is that it does not guarantee a *sufficient decrease* in the values of the function f , a notion that we discuss next.

5.1.2.1 The Wolfe Conditions

Definition 5.2 (*Sufficient decrease*) We say that a stepsize α provides a *sufficient decrease* of f if

$$f(x_k + \alpha p_k) \leq f(x_k) + c_1 \alpha \nabla f_k^T p_k, \quad (5.7)$$

where c_1 is a constant in $]0, 1[$. Inequality (5.7) is also known as the *Armijo rule*.

In terms of the function ϕ defined in (5.6), this condition is equivalent to

$$\phi(\alpha) \leq \phi(0) + c_1 \alpha \phi'(0).$$

The linear function

$$l(\alpha) := c_1 \phi'(0) \alpha + \phi(0)$$

has negative slope $c_1 \nabla f_k^T p_k$, but lies above the graph of $\phi(\alpha)$ for small values of α , as a consequence of $c_1 \in]0, 1[$. The sufficient decrease condition establishes that α is acceptable if and only if

$$\phi(\alpha) \leq l(\alpha);$$

See Fig. 5.5. In practice, c_1 is usually chosen to be quite small. Nocedal and Wright [70] propose a value of $c_1 = 10^{-4}$.

Although this first rule avoids situations like the one showed in Fig. 5.4, it is satisfied with very small values of α . If these values were taken for α_k , the algorithm

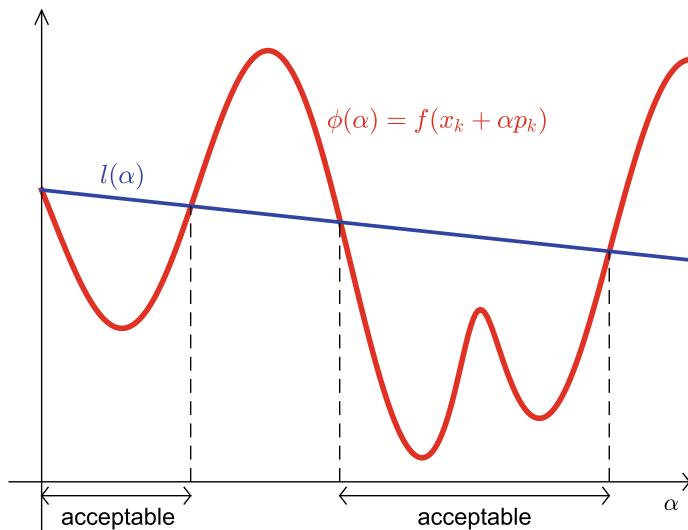


Fig. 5.5 Sufficient decrease condition

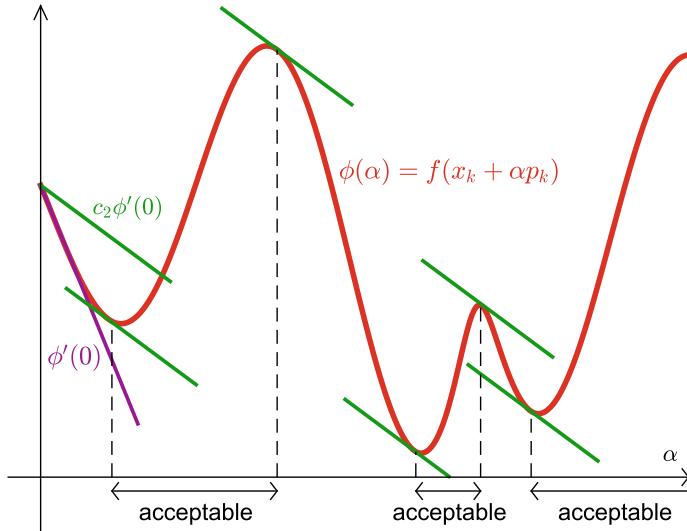


Fig. 5.6 Curvature condition

would not provide a *reasonable* progress. In order to exclude *excessively short* steps, the following additional condition is introduced.

Definition 5.3 (*Curvature condition*) The *curvature condition* requires that α_k satisfies

$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k, \quad (5.8)$$

for some constant $c_2 \in]c_1, 1[$.

In terms of the function ϕ , condition (5.8) is equivalent to

$$\phi'(\alpha_k) \geq c_2\phi'(0);$$

that is, the curvature condition ensures that the slope of the curve ϕ in α_k is greater or equal to c_2 times the slope of ϕ in 0. This is a reasonable condition to be required: If the slope $\phi'(\alpha)$ is smaller than $c_2\phi'(0)$, this indicates that we could significantly reduce f by moving farther along the chosen search direction. The curvature condition is illustrated in Fig. 5.6. Nocedal and Wright [70] give an example value for c_2 of 0.9 when p_k is obtained by Newton or quasi-Newton methods, and of 0.1 when p_k is obtained by the conjugate gradient method.

Definition 5.4 (*Wolfe conditions*) The sufficient decrease and the curvature conditions are known together as the *Wolfe conditions*:

$$f(x_k + \alpha_k p_k) \leq f(x_k) + c_1 \alpha_k \nabla f_k^T p_k, \quad (5.9)$$

$$\nabla f(x_k + \alpha_k p_k)^T p_k \geq c_2 \nabla f_k^T p_k, \quad (5.10)$$

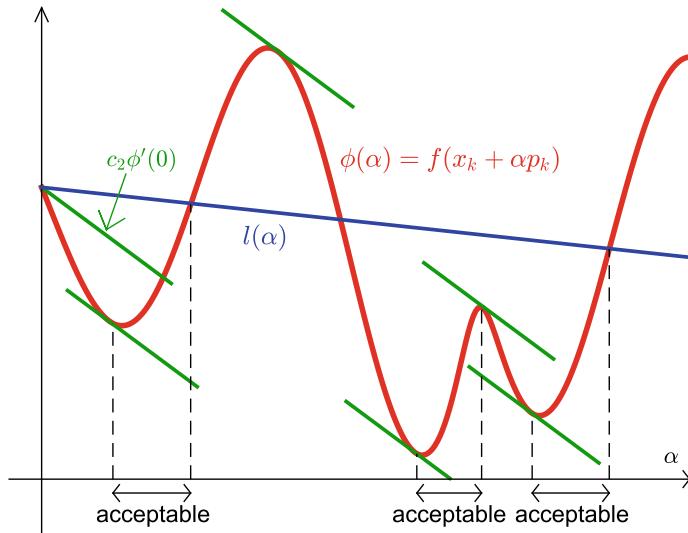


Fig. 5.7 Stepsizes that satisfy the Wolfe conditions

where $0 < c_1 < c_2 < 1$.

A stepsize can satisfy the Wolfe conditions without being particularly close to a minimum of ϕ , as it is shown in Fig. 5.7.

As we see next, it is not difficult to find stepsizes that satisfy the Wolfe conditions for every function f which is smooth and bounded below. This can be done by using the following algorithm, which employs a bisection technique.

Algorithm 1: Bisection algorithm for the Wolfe conditions

```

Choose any  $0 < c_1 < c_2 < 1$ . Set  $a = 0$ ,  $b = 2$ ,  $continue = \text{true}$ .
while  $f(x + bp) \leq f(x) + c_1 bf'(x; p)$  do
|  $b = 2b$ 
end
while  $continue$  do
|  $\alpha = \frac{1}{2}(a + b)$ 
| if  $f(x + \alpha p) > f(x) + c_1 \alpha f'(x; p)$  then
| |  $b = \alpha$ 
| else if  $f'(x + \alpha p; p) < c_2 f'(x; p)$  then
| |  $a = \alpha$ 
| else
| |  $continue = \text{false}$ 
| end
end
```

Proposition 5.5 Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable. Let p_k be a descent direction at x_k and assume that f is bounded below along the half-line $\{x_k + \alpha p_k : \alpha > 0\}$. If $0 < c_1 < c_2 < 1$, then Algorithm 1 returns in a finite number of iterations a stepsize satisfying the Wolfe conditions (5.9) and (5.10).

Proof Since f is bounded below along the half-line $\{x + \alpha p \mid \alpha > 0\}$, the first **while** loop must end after a finite number of iterations. Thus, there exists some $b_0 \geq 2$ such that

$$f(x + b_0 p) > f(x) + c_1 b_0 f'(x; p).$$

Take $a_0 := 0$, and let us denote by $a_k < \alpha_k < b_k$ the values of a , b , and α at the iteration $k = 1, 2, \dots$ of the second **while** loop. Reasoning by contradiction, suppose that this loop is infinite. Observe that, for all $k \in \mathbb{N}$, the following properties hold:

- a_k satisfies the Armijo condition; i.e.,

$$f(x + a_k p) \leq f(x) + c_1 a_k f'(x; p). \quad (5.11)$$

- b_k does not satisfy the Armijo condition; i.e.,

$$f(x + b_k p) > f(x) + c_1 b_k f'(x; p). \quad (5.12)$$

- a_k does not satisfy the curvature condition; i.e.,

$$f'(x + a_k p; p) < c_2 f'(x; p). \quad (5.13)$$

Moreover, we have that $b_{k+1} - a_{k+1} = \frac{1}{2}(b_k - a_k)$ for all $k \in \mathbb{N}$:

- If $a_{k+1} = a_k$ and $b_{k+1} = \alpha_k$, then $b_{k+1} - a_{k+1} = \alpha_k - a_k = \frac{1}{2}(b_k - a_k)$.
- If $a_{k+1} = \alpha_k$ and $b_{k+1} = b_k$, then $b_{k+1} - a_{k+1} = b_k - \alpha_k = \frac{1}{2}(b_k - a_k)$.

Therefore, since $a_k < \alpha_k < b_k$, there exists some α^* such that

$$a_k \nearrow \alpha^*, \quad \alpha_k \rightarrow \alpha^*, \quad b_k \searrow \alpha^*.$$

Taking limits in (5.13) as $k \rightarrow \infty$, we obtain

$$f'(x + \alpha^* p; p) \leq c_2 f'(x; p). \quad (5.14)$$

Subtracting (5.12) to (5.11) and applying the mean value theorem, we deduce that there exists $\widehat{\alpha}_k \in]a_k, b_k[$ such that

$$c_1(b_k - a_k) f'(x; p) < f(x + b_k p) - f(x + a_k p) = (b_k - a_k) f'(x + \widehat{\alpha}_k p; p),$$

Dividing the latter expression by $(b_k - a_k) > 0$, taking limits when $k \rightarrow \infty$, and using (5.14), we deduce

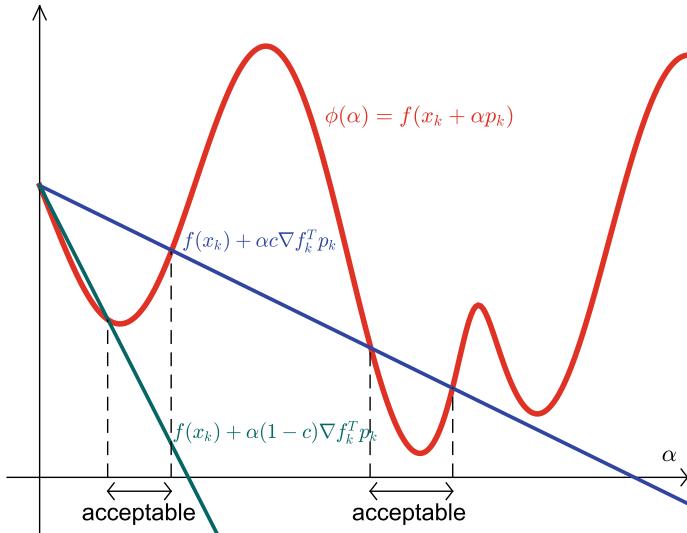


Fig. 5.8 Goldstein conditions

$$c_1 f'(x; p) \leq f'(x + \alpha^* p; p) \leq c_2 f'(x; p).$$

This is a contradiction, because $0 < c_1 < c_2$ and $f'(x; p) < 0$. \square

5.1.2.2 Goldstein and Backtracking Conditions

The *Goldstein conditions* also ensure the achievement of a sufficient decrease, while avoiding at the same time stepsizes α that are too small, acting thus in a similar way than Wolfe conditions (5.9) and (5.10). They are defined by the next two inequalities:

$$f(x_k) + (1 - c)\nabla f_k^T p_k \leq f(x_k + \alpha_k p_k) \leq f(x_k) + c\alpha_k \nabla f_k^T p_k, \quad (5.15)$$

where $0 < c < 1/2$. The second inequality is just the sufficient decrease condition (5.7), while the first inequality is introduced to avoid small stepsizes to be chosen (see Fig. 5.8).

A disadvantage of the Goldstein conditions compared with Wolfe conditions is that the first inequality in (5.15) can exclude some minima of ϕ . Nevertheless, the Wolfe and Goldstein conditions have much in common, and their convergence results are quite similar.

We have seen that the sufficient decrease condition (5.9) is not enough to guarantee a reasonable decrease along the chosen search direction. Nonetheless, if the line search algorithm chooses the stepsize candidates by the so-called *backtracking* procedure, we can omit condition (5.10) and use only the sufficient decrease condition.

In its most basic form, it consists in the following: Given two constants $c, \rho \in]0, 1[$, the backtracking procedure starts from an initial point $\alpha = \bar{\alpha} > 0$ at which condition (5.9) is verified. If it is not fulfilled, one takes $\rho\alpha$ as the new value of α and the procedure is repeated until this condition is achieved:

Algorithm 2: Backtracking

```

Choose  $\bar{\alpha} > 0, \rho, c \in ]0, 1[$ . Take  $\alpha = \bar{\alpha}$ .
while  $f(x_k + \alpha p_k) > f(x_k) + c\alpha \nabla f_k^T p_k$  do
|  $\alpha = \rho\alpha$ ;
end
return  $\alpha_k = \alpha$ 
```

With the backtracking procedure, one ensures that either the stepsize takes a fixed value (the initial $\bar{\alpha}$), or it satisfies the sufficient decrease condition while trying to avoid picking very small values.

5.2 Convergence of the Line Search Methods

To obtain the *global convergence* of an algorithm, that is, the convergence for any starting point x_0 , we should not only make an appropriate selection of the stepsizes, but also an adequate choice of the search directions p_k . In this section, we will focus on the requirements about the search directions, paying attention to a key element: the angle θ_k between p_k and the steepest descent direction $-\nabla f_k$, defined by

$$\cos \theta_k = \frac{-\nabla f_k^T p_k}{\|\nabla f_k\| \|p_k\|}. \quad (5.16)$$

Next theorem will let us conclude that the steepest descent method is globally convergent. For any other algorithm, it describes how much p_k can deviate from the steepest descent method while keeping the global convergence guaranteed.

Theorem 5.6 (Zoutendijk's theorem) *Consider an algorithm of the form $x_{k+1} = x_k + \alpha_k p_k$, where p_k is a descent direction and α_k satisfies the Wolfe conditions (5.9) and (5.10). Suppose that f is bounded below on \mathbb{R}^n and that $f \in C^1(U)$, with U being an open set that contains the sublevel set $S_{f(x_0)}(f) = \{x \in \mathbb{R}^n : f(x) \leq f(x_0)\}$, where x_0 is an initial point. Assume that ∇f is Lipschitz continuous on U ; i.e., there exists some $\lambda > 0$ such that*

$$\|\nabla f(x) - \nabla f(y)\| \leq \lambda \|x - y\|, \quad \forall x, y \in U.$$

Then, one has

$$\sum_{k=0}^{\infty} (\cos^2 \theta_k) \|\nabla f(x_k)\|^2 < \infty. \quad (5.17)$$

Proof By the second Wolfe condition (5.10) and since $x_{k+1} = x_k + \alpha_k p_k$, we have

$$(\nabla f_{k+1} - \nabla f_k)^T p_k \geq (c_2 - 1) \nabla f_k^T p_k \geq 0.$$

Further, using the Lipschitz condition, we get

$$(\nabla f_{k+1} - \nabla f_k)^T p_k \leq \|\nabla f_{k+1} - \nabla f_k\| \|p_k\| \leq \lambda \alpha_k \|p_k\|^2.$$

Combining these two relationships, we obtain

$$\alpha_k \geq \left(\frac{c_2 - 1}{\lambda} \right) \frac{\nabla f_k^T p_k}{\|p_k\|^2}.$$

By substituting this inequality in the first Wolfe condition (5.9), we deduce

$$f_{k+1} \leq f_k - (-\alpha_k) c_1 \nabla f_k^T p_k \leq f_k - c_1 \left(\frac{1 - c_2}{\lambda} \right) \frac{(\nabla f_k^T p_k)^2}{\|p_k\|^2}.$$

Using (5.16), we can express this relationship as

$$f_{k+1} \leq f_k - c \cos^2 \theta_k \|\nabla f_k\|^2,$$

where $c = c_1(1 - c_2)/\lambda$. Adding this expression for every index smaller or equal to k , we obtain

$$f_{k+1} \leq f_0 - c \sum_{j=0}^k \cos^2 \theta_j \|\nabla f_j\|^2. \quad (5.18)$$

As f is bounded below, we have that $f_0 - f_{k+1}$ is smaller than some given positive constant, for every $k \in \mathbb{N}$. Taking limits in (5.18), we deduce (5.17). \square

Similar results can be obtained when the Goldstein conditions (5.15) are used instead of the Wolfe conditions.

Observe that the hypotheses in the previous theorem are not excessively restrictive. If the function f was not bounded below, the optimization problem would not be considered as *well defined*. The *smoothness* hypothesis (or Lipschitz continuity of the gradient) is implied by many of the local convergence conditions of the most representative algorithms.

Proposition 5.7 *Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be such that $f \in C^2(U)$, where U is an open convex set, and suppose that $\nabla^2 f$ is bounded on U . Then, ∇f is Lipschitz continuous on U .*

Proof For every $x, y \in U$, one has

$$\nabla f(y) - \nabla f(x) = \int_0^1 \nabla^2 f(x + t(y - x))(y - x) dt.$$

Taking norms, we obtain

$$\begin{aligned}\|\nabla f(y) - \nabla f(x)\| &\leq \int_0^1 \|\nabla^2 f(x + t(y-x))(y-x)\| dt \\ &\leq \int_0^1 \|\nabla^2 f(x + t(y-x))\| \|y-x\| dt.\end{aligned}$$

As $\nabla^2 f$ is bounded on U , there exists some constant $\lambda > 0$ such that $\|\nabla^2 f(z)\| \leq \lambda$, for every $z \in U$. Since U is convex, if $t \in [0, 1]$, one has that $x + t(y-x) \in U$. Hence $\|\nabla^2 f(x + t(y-x))\| \leq \lambda$ and we deduce that

$$\|\nabla f(y) - \nabla f(x)\| \leq \int_0^1 \lambda \|x-y\| dt = \lambda \|x-y\|,$$

for every $x, y \in U$. \square

Remark 5.8 In the hypotheses of Theorem 5.6, we only require ∇f to be Lipschitz continuous on U , not on the entire space. For instance, for the function $f(x) = x^4$, one has that

$$|f'(x) - f'(y)| = 4|x^3 - y^3| = 4|x^2 + xy + y^2| |x - y|.$$

The expression $4|x^2 + xy + y^2| = 3(x+y)^2 + (x-y)^2$ is not bounded on the real line; nonetheless, it is bounded on every bounded set U ; see Fig. 5.9. Therefore, f' is Lipschitz continuous on every bounded set U , but not on the real line.

The property (5.17), known as *Zoutendijk's condition*, implies that

$$\cos^2 \theta_k \|\nabla f(x_k)\|^2 \rightarrow 0. \quad (5.19)$$

This limit can be used to deduce global convergence results for line search algorithms. If the method for the selection of p_k ensures that the angle θ_k is bounded above by θ , and this bound is smaller than $\pi/2$, there exists a positive constant δ such that

$$\cos \theta_k \geq \cos \theta = \delta > 0, \quad \forall k \in \mathbb{N}.$$

It follows then from (5.19) that

$$\lim_{k \rightarrow \infty} \|\nabla f(x_k)\| = 0. \quad (5.20)$$

In other words, we can ensure that $\nabla f(x_k) \rightarrow 0_n$ as long as the search directions are kept *uniformly* apart from orthogonality with respect to the gradient. In particular, the steepest descent method (for which $\theta_k = 0$ for all k) trivially satisfies this condition. Therefore, it generates a sequence x_k such that $\nabla f(x_k)$ converges to 0_n when $k \rightarrow \infty$, whenever the line searches satisfy the Wolfe conditions (5.9) and (5.10).

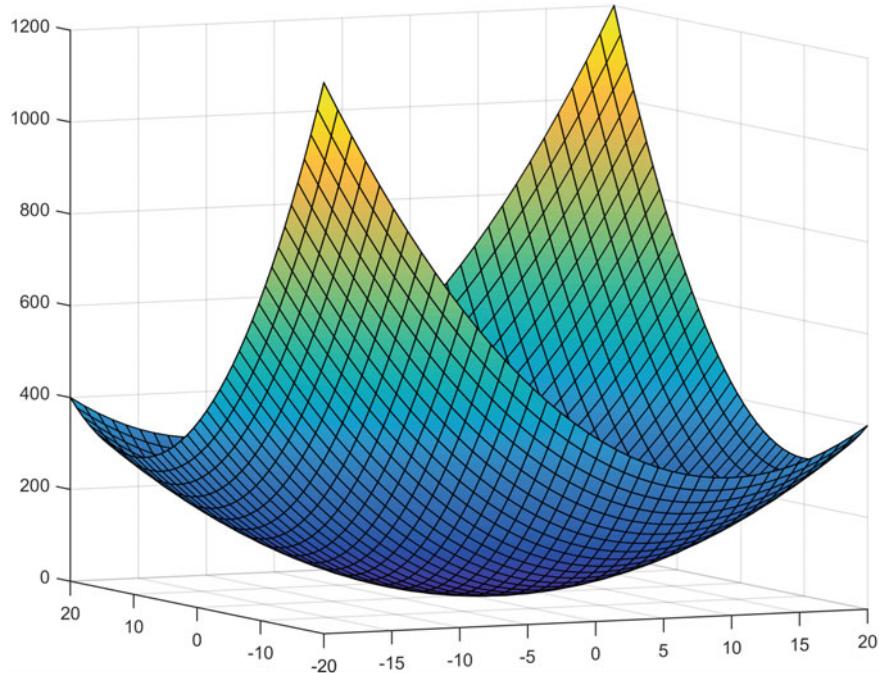


Fig. 5.9 The function $3(x + y)^2 + (x - y)^2$ is bounded on every bounded set U , but it is not bounded on $\mathbb{R} \times \mathbb{R}$

Property (5.20) constitutes a prototypical result of global convergence, as the validity of the result does not depend on the position of the initial point x_0 . It is important to remark that this result does not guarantee that the method converges to a local minimum, not even to a stationary point. Nevertheless, observe that if $S_{f(x_0)}(f)$ is bounded, since $\{x_k\} \subset S_{f(x_0)}(f)$, there exists a subsequence which is convergent to a point $\bar{x} \in S_{f(x_0)}(f)$. To simplify, let us assume without loss of generality that the entire sequence $\{x_k\}$ converges to \bar{x} . Since $f \in C^1(U)$ and $S_{f(x_0)}(f) \subset U$, we have

$$\nabla f(\bar{x}) = \nabla f\left(\lim_{k \rightarrow \infty} x_k\right) = \lim_{k \rightarrow \infty} \nabla f(x_k) = 0_n,$$

and we deduce that \bar{x} is a stationary point.

Proposition 5.9 *Consider a gradient method of the form*

$$p_k = -B_k^{-1} \nabla f_k,$$

where B_k are positive definite matrices with a conditioning number uniformly bounded by some constant $M > 0$, that is,

$$\text{cond}(B_k) = \|B_k\| \|B_k^{-1}\| \leq M, \quad \forall k \in \mathbb{N}.$$

Then,

$$\cos \theta_k \geq \frac{1}{M}, \quad \forall k \in \mathbb{N},$$

and therefore, under the same hypotheses of Theorem 5.6, one has $\nabla f_k \rightarrow 0_n$.

Proof Indeed, if $\lambda_{\min}(B_k)$ and $\lambda_{\max}(B_k)$ are the smallest and the largest eigenvalue of B_k , respectively, one has

$$\begin{aligned} \cos \theta_k &= -\frac{\nabla f_k^T p_k}{\|\nabla f_k\| \|p_k\|} = \frac{\nabla f_k^T B_k^{-1} \nabla f_k}{\|\nabla f_k\| \|B_k^{-1} \nabla f_k\|} \geq \frac{\|\nabla f_k\|^2 \frac{1}{\lambda_{\max}(B_k)}}{\|\nabla f_k\| \|B_k^{-1} \nabla f_k\|} \\ &\geq \frac{\|\nabla f_k\| \frac{1}{\lambda_{\max}(B_k)}}{\|B_k^{-1}\| \|\nabla f_k\|} = \frac{1}{\frac{\lambda_{\max}(B_k)}{\lambda_{\min}(B_k)}} = \frac{1}{\text{cond}(B_k)} \geq \frac{1}{M}, \end{aligned}$$

where in the first inequality we have used the fact that for every symmetric matrix A , one has that $\lambda_{\min}(A)\|z\|^2 \leq z^T Az \leq \lambda_{\max}(A)\|z\|^2$; see (1.12). \square

5.2.1 Rate of Convergence

The mere fact that a sequence $\{x_k\}$ converges to a stationary point \bar{x} may have a relative interest in practice unless the *rate of convergence* is satisfactory.

There are different criteria for quantifying the rate of convergence of an algorithm. One could study the *computational complexity* of the algorithm, either by estimating the number of elementary operations needed to find an exact or approximate solution or by analyzing the number of function evaluations (and possibly, gradient evaluations) carried out during the application of the algorithm. The problem with this second methodology is that it considers the *worst case*, and practice shows that high computational complexity algorithms have better average performance than others with less complexity. This occurs because the cases where the first algorithms behave badly are unlikely in real situations.

The next discussion is focused on the local analysis of the algorithm, whose main features are the following:

- It is confined to sequences $\{x_k\}$ converging to a limit point \bar{x} .
- The *rate of convergence* is estimated by means of an *error function* $e : \mathbb{R}^n \rightarrow \mathbb{R}_+$, which is continuous and verifies $e(\bar{x}) = 0$. Some usual error functions are:
 - $e(x) = \|x - \bar{x}\|$ (Euclidean distance);
 - $e(x) = |f(x) - f(\bar{x})|$ (absolute error in the function value).
- Our analysis is asymptotic; that is, we look at the rate of convergence of the tail of the sequence of errors $\{e(x_k)\}$.

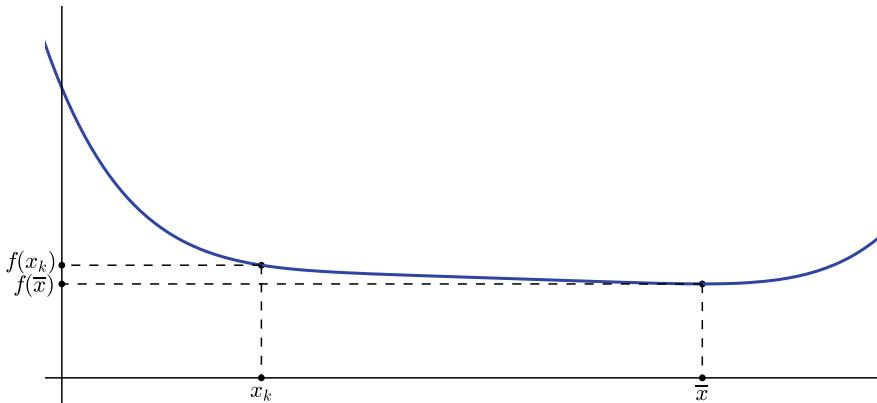


Fig. 5.10 x_k is far from \bar{x} despite that $f(x_k)$ is close to $f(\bar{x})$

We would like to know either how fast x_k converges to \bar{x} , or how fast $f(x_k)$ approaches $f(\bar{x})$. It may occur that we are approaching fast the value of the function $f(\bar{x})$ and quite slowly to the point \bar{x} , situation illustrated in Fig. 5.10.

Definition 5.10 We say that $\{e(x_k)\}$ converges linearly if there is a constant $\beta \in]0, 1[$ such that

$$\limsup_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} = \lim_{k \rightarrow \infty} \left(\sup_{m \geq k} \frac{e(x_{m+1})}{e(x_m)} \right) \leq \beta. \quad (5.21)$$

When the latter inequality is valid for all $\beta \in]0, 1[, i.e., if$

$$\lim_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} = 0,$$

we say that $\{e(x_k)\}$ converges superlinearly. If the inequality (5.21) is not verified for any $\beta \in]0, 1[, we say that $\{e(x_k)\}$ converges sublinearly.$

A sequence that converges sublinearly is considered in practice as not convergent, as the convergence may be so slow that an algorithm with this rate should not be used.

The following concept yields a refinement of the notion of superlinear convergence.

Definition 5.11 We say that $\{e(x_k)\}$ converges superlinearly with order p , for $p > 1$, when

$$\limsup_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)^p} < \infty. \quad (5.22)$$

The case $p = 2$ is known as quadratic convergence.

Let us introduce our last type of convergence.

Definition 5.12 We say that $\{e(x_k)\}$ converges geometrically when there exist some constants $q > 0$ and $\gamma \in]0, 1[$ such that

$$e(x_k) \leq q\gamma^k, \quad \forall k. \quad (5.23)$$

Proposition 5.13 Linear convergence implies geometric convergence.

Proof Given $\beta \in]0, 1[$ satisfying (5.21), if we take $\gamma \in]\beta, 1[$, there must exist k_0 such that

$$\frac{e(x_{k+1})}{e(x_k)} \leq \gamma, \quad \forall k \geq k_0,$$

and from this inequality we get

$$e(x_{k_0+p}) \leq \gamma^p e(x_{k_0}), \quad \forall p \geq 1. \quad (5.24)$$

Defining $q := \max\{e(x_k)/\gamma^k, k = 1, 2, \dots, k_0\}$, we can write

$$e(x_k) \leq q\gamma^k, \quad k = 1, 2, \dots, k_0,$$

and replacing this in (5.24),

$$e(x_{k_0+p}) \leq \gamma^p e(x_{k_0}) \leq q\gamma^{k_0+p}, \quad \forall p \geq 1,$$

and therefore, we conclude that (5.23) holds. \square

Remark 5.14 The converse implication is not true: geometric convergence does not imply linear convergence. As an example, let us consider $e(x_{2p}) = \gamma^{3p+1}$ and $e(x_{2p+1}) = \gamma^{2p+1}$, with $\gamma \in]0, 1[$. One has $e(x_k) \leq \gamma^k$, but

$$\limsup_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} = \lim_{p \rightarrow \infty} \frac{e(x_{2p+1})}{e(x_{2p})} = \lim_{p \rightarrow \infty} \frac{\gamma^{2p+1}}{\gamma^{3p+1}} = \lim_{p \rightarrow \infty} \frac{1}{\gamma^{p+1}} = \infty.$$

Hence, $\{e(x_k)\}$ does not converge linearly.

The relation $e(x_{k+1}) \leq \gamma e(x_k)$, for $k \geq k_0$, means that, asymptotically, the error is reduced in each iteration by a factor that is at least γ . This is the reason why it is called linear convergence, because the errors $e(x_k)$ remain below the line $y = \gamma x$ (see Fig. 5.11).

According to the definition of \limsup , and using the big O notation, it turns out that (5.22) is equivalent to

$$e(x_{k+1}) = O(e(x_k)^p),$$

i.e., to the existence of $q > 0$ and $k_0 \geq 0$ such that $e(x_{k+1}) \leq q e(x_k)^p$, for all $k \geq k_0$. From the last inequality we obtain the interpretation shown in Fig. 5.12.

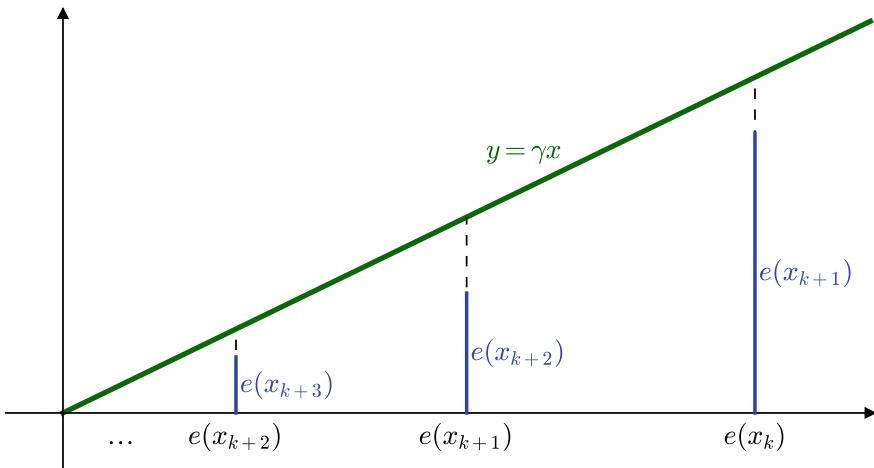


Fig. 5.11 Linear convergence

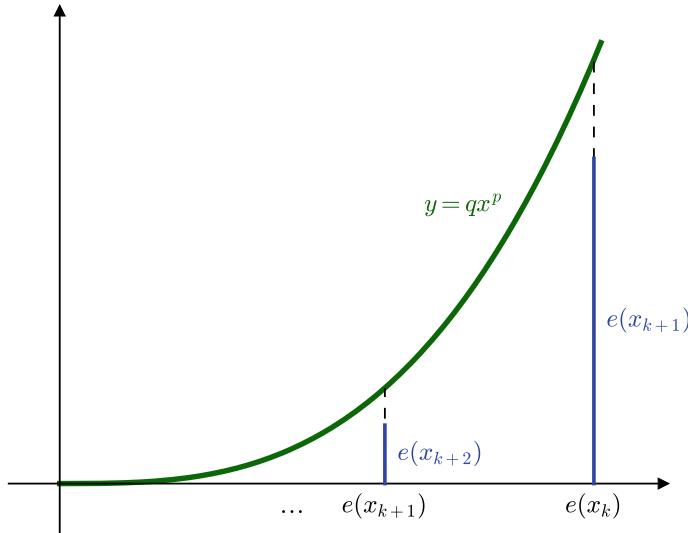


Fig. 5.12 Superlinear convergence with order $p > 1$

Proposition 5.15 *Superlinear convergence with order p implies superlinear convergence.*

Proof Let us suppose that

$$\limsup_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)^p} < M,$$

for a certain $M > 0$. Then, there exists k_0 such that

$$\frac{e(x_{k+1})}{e(x_k)^p} \leq M, \quad \forall k \geq k_0,$$

or, equivalently,

$$\frac{e(x_{k+1})}{e(x_k)} \leq M e(x_k)^{p-1}, \quad \forall k \geq k_0.$$

Thus, for all $n \geq k_0$,

$$\sup_{k \geq n} \frac{e(x_{k+1})}{e(x_k)} \leq \sup_{k \geq n} M e(x_k)^{p-1}.$$

Since $p > 1$ and $e(x_k)$ converges to zero, taking limits as $n \rightarrow \infty$ in the last inequality, we get

$$\begin{aligned} \limsup_{k \rightarrow \infty} \frac{e(x_{k+1})}{e(x_k)} &= \limsup_{n \rightarrow \infty} \sup_{k \geq n} \frac{e(x_{k+1})}{e(x_k)} \leq \limsup_{n \rightarrow \infty} \sup_{k \geq n} M e(x_k)^{p-1} \\ &= \lim_{k \rightarrow \infty} M e(x_k)^{p-1} = 0. \end{aligned}$$

Hence, $\lim_{k \rightarrow \infty} e(x_{k+1})/e(x_k) = 0$. □

We have thus proved the following chain of implications between the rates of convergence:

superlinear with order $p > 1 \Rightarrow$ superlinear \Rightarrow linear \Rightarrow geometric.

5.2.2 Convergence Rate of Gradient Methods for Quadratic Forms

In order to study the rate of convergence of the gradient methods, we analyze the simplest situation corresponding to the case where the objective function is quadratic. If the function is not quadratic but is twice continuously differentiable and \bar{x} is a local minimum, by Taylor's theorem, the function f can be approximated around \bar{x} by the quadratic function

$$f(\bar{x}) + \frac{1}{2}(x - \bar{x})^T \nabla^2 f(\bar{x})(x - \bar{x}),$$

so we may expect the asymptotic convergence results obtained for the quadratic case to be similar to those of the general case. We therefore shall assume that f is a quadratic function whose Hessian matrix Q is symmetric and positive definite. Without loss of generality, we can even suppose that f has its global minimum at $\bar{x} = 0_n$ and that $f(\bar{x}) = 0$. Indeed, if $f(x) = \frac{1}{2}x^T Qx - c^T x$, the minimum \bar{x} of f is the unique solution of the system $Q\bar{x} = c$. Then, \bar{x} is also the global minimum of the function $h(x) = \frac{1}{2}(x - \bar{x})^T Q(x - \bar{x}) = f(x) - f(\bar{x})$. Accordingly,

$$f(x) = \frac{1}{2}x^T Qx, \quad \nabla f(x) = Qx, \quad \nabla^2 f(x) = Q. \quad (5.25)$$

Method of the Steepest Descent

The steepest descent method applied to the quadratic function (5.25) generates the iterates

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) = (I - \alpha_k Q)x_k.$$

Hence,

$$\|x_{k+1}\|^2 = x_k^T (I - \alpha_k Q)^2 x_k \leq \lambda_{\max}((I - \alpha_k Q)^2) \|x_k\|^2.$$

We know that the eigenvalues of the matrix $(I - \alpha_k Q)^2$ are $(1 - \alpha_k \lambda_i)^2$, where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of the matrix Q . If we represent by $\lambda_{\min} = \lambda_{\min}(Q)$ and $\lambda_{\max} = \lambda_{\max}(Q)$, we obviously have

$$\lambda_{\max}((I - \alpha_k Q)^2) = \max\{(1 - \alpha_k \lambda_{\min})^2, (1 - \alpha_k \lambda_{\max})^2\}.$$

It follows that, for $x_k \neq 0_n$,

$$\frac{\|x_{k+1}\|}{\|x_k\|} \leq \max\{|1 - \alpha_k \lambda_{\min}|, |1 - \alpha_k \lambda_{\max}|\}. \quad (5.26)$$

From Fig. 5.13, we see that the value of α_k minimizing this upper bound is

$$\bar{\alpha}_k = \frac{2}{\lambda_{\max} + \lambda_{\min}}.$$

In this case,

$$\frac{\|x_{k+1}\|}{\|x_k\|} \leq \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\frac{\lambda_{\max}}{\lambda_{\min}} - 1}{\frac{\lambda_{\max}}{\lambda_{\min}} + 1} = \frac{\text{cond}(Q) - 1}{\text{cond}(Q) + 1}. \quad (5.27)$$

This is the best upper bound for the convergence rate of the steepest descent method when applied to the function (5.25) and the stepsize is constant. Let us observe that, thanks to (5.26), the convergence is guaranteed for any stepsize α_k satisfying

$$\max\{|1 - \alpha_k \lambda_{\min}|, |1 - \alpha_k \lambda_{\max}|\} < 1,$$

in other words, for all $\alpha_k \in]0, 2/\lambda_{\max}[$ (see Fig. 5.13).

Another result of interest about the rate of convergence of the steepest descent method is derived for the case when α_k is chosen by an exact line search. This result quantifies the rate at which the cost function drops:

$$\frac{f(x_{k+1})}{f(x_k)} \leq \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 = \left(\frac{\text{cond}(Q) - 1}{\text{cond}(Q) + 1} \right)^2. \quad (5.28)$$

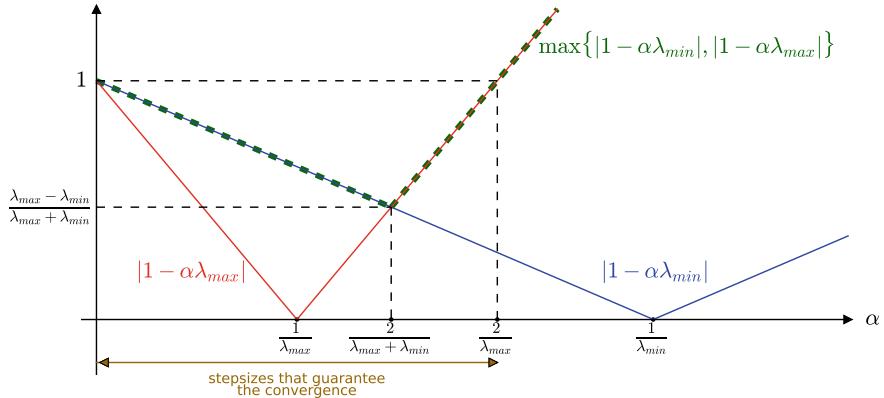


Fig. 5.13 The upper bound is minimized when $1 - \alpha\lambda_{\min} = \alpha\lambda_{\max} - 1$, i.e., at $\bar{\alpha} = \frac{2}{\lambda_{\max} + \lambda_{\min}}$

From (5.27) and (5.28), it follows that the steepest descent method may converge very slowly when the condition number of Q is large; see Fig. 5.14. Conversely, if $\text{cond}(Q) \approx 1$, the convergence will be good. Compare Fig. 5.14 with Fig. 5.15. In the best case, when $\text{cond}(Q) = 1$, the method provides the optimum in one step. Note that, as far as (5.27) and (5.28) are less than 1, the rate of convergence is linear. In Fig. 5.14, $Q = \begin{bmatrix} 1 & 0 \\ 0 & 1000 \end{bmatrix}$ and so, $\text{cond}(Q) = 1000$. In the figure, we show the level curves of the function $f(x) = \frac{1}{2}x^T Qx$ and the iterates (in red) produced by the steepest descent method if $x_0 = (1, 1/1000)^T$. The convergence is very slow. (We are taking different scales in both axes.)

In the example illustrated in Fig. 5.15, $Q = \begin{bmatrix} 1 & 0 \\ 0 & 10 \end{bmatrix}$, and thus, $\text{cond}(Q) = 10$. We show the level curves of the function $f(x) = \frac{1}{2}x^T Qx$ and the steepest descent method iterates for $x_0 = (1, 1/1000)^T$. The process converges slowly and zigzagging but faster than in Fig. 5.14.

In the proof of (5.28), we shall use the following result.

Lemma 5.16 (Kantorovich's inequality) *Let Q be an $n \times n$ symmetric and positive definite matrix. Then, the following inequality holds for all $y \neq 0_n$*

$$\frac{(y^T y)^2}{(y^T Q y)(y^T Q^{-1} y)} \geq \frac{4\lambda_{\max}\lambda_{\min}}{(\lambda_{\max} + \lambda_{\min})^2}, \quad (5.29)$$

where λ_{\min} and λ_{\max} are the smallest and the largest eigenvalue of Q , respectively.

Proof Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of Q , and assume that

$$0 < \lambda_{\min} = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n = \lambda_{\max}.$$

Let U be an orthogonal matrix such that $U^T Q U = \text{diag}(\lambda_1, \dots, \lambda_n)$. Accordingly, we can suppose, without loss of generality (making the coordinate transformation

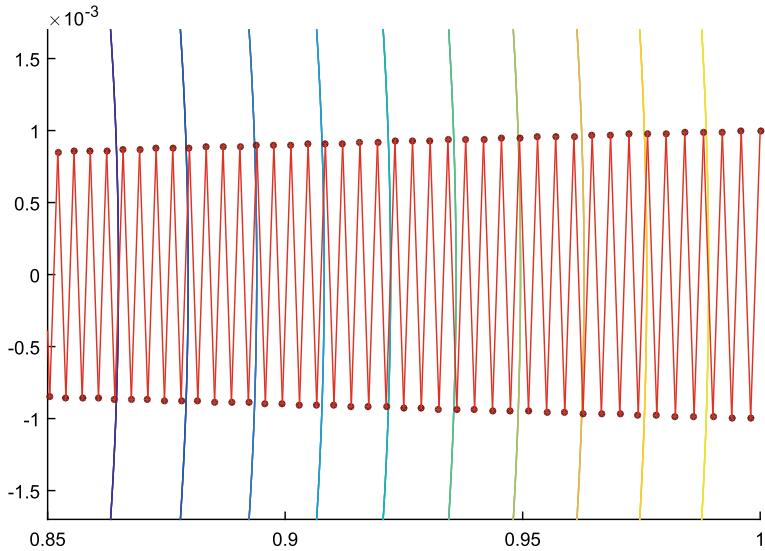
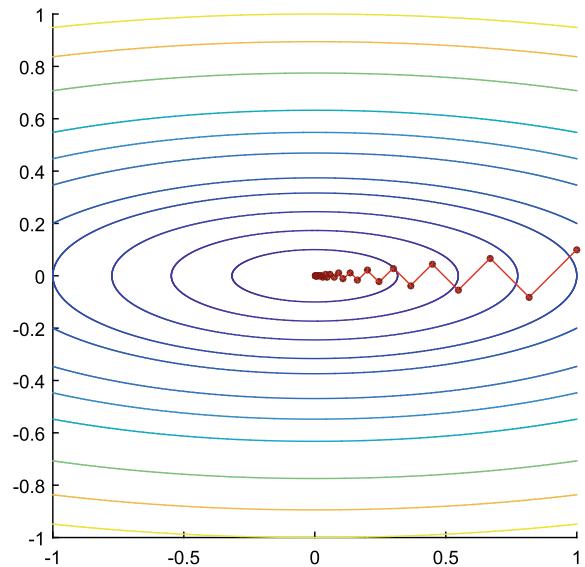


Fig. 5.14 Slow convergence of the steepest descent method

Fig. 5.15 The steepest descent method converges slowly and zigzagging but faster than in Fig. 5.14



which replaces y by Ux), that Q is the diagonal matrix $\text{diag}(\lambda_1, \dots, \lambda_n)$. Thus, for all $y = (y_1, \dots, y_n)^T \neq 0_n$,

$$\frac{(y^T y)^2}{(y^T Q y)(y^T Q^{-1} y)} = \frac{\left(\sum_{i=1}^n y_i^2\right)^2}{\left(\sum_{i=1}^n \lambda_i y_i^2\right) \left(\sum_{i=1}^n \frac{y_i^2}{\lambda_i}\right)}.$$

Let us consider now the convex function $\phi(\lambda) = 1/\lambda$ and take $\xi = (\xi_1, \dots, \xi_n)^T$ with

$$\xi_j := \frac{y_j^2}{\sum_{i=1}^n y_i^2}, \quad j = 1, \dots, n.$$

Then, we have

$$\frac{(y^T y)^2}{(y^T Q y)(y^T Q^{-1} y)} = \frac{1}{\left(\sum_{i=1}^n \lambda_i \xi_i\right)\left(\sum_{i=1}^n \phi(\lambda_i) \xi_i\right)}. \quad (5.30)$$

Now, we introduce

$$\lambda := \sum_{i=1}^n \lambda_i \xi_i \quad \text{and} \quad \lambda_\phi := \sum_{i=1}^n \phi(\lambda_i) \xi_i.$$

Since $\xi_i \geq 0$ and $\sum_{i=1}^n \xi_i = 1$, it is obvious that $\lambda_{min} \leq \lambda \leq \lambda_{max}$. Suppose that $\lambda_{min} \neq \lambda_{max}$ (otherwise, the ratio in (5.30) is equal to 1 and (5.29) holds with equality). Each λ_i can be represented as a convex linear combination of λ_{min} and λ_{max} :

$$\lambda_i = \frac{\lambda_i - \lambda_{max}}{\lambda_{min} - \lambda_{max}} \lambda_{min} + \frac{\lambda_{min} - \lambda_i}{\lambda_{min} - \lambda_{max}} \lambda_{max}.$$

The convexity of ϕ gives rise to

$$\phi(\lambda_i) \leq \frac{\lambda_i - \lambda_{max}}{\lambda_{min} - \lambda_{max}} \phi(\lambda_{min}) + \frac{\lambda_{min} - \lambda_i}{\lambda_{min} - \lambda_{max}} \phi(\lambda_{max}).$$

Therefore,

$$\begin{aligned} \lambda_\phi &\leq \sum_{i=1}^n \left(\frac{\lambda_i - \lambda_{max}}{\lambda_{min} - \lambda_{max}} \phi(\lambda_{min}) + \frac{\lambda_{min} - \lambda_i}{\lambda_{min} - \lambda_{max}} \phi(\lambda_{max}) \right) \xi_i \\ &= \sum_{i=1}^n \frac{\lambda_{min} + \lambda_{max} - \lambda_i}{\lambda_{min} \lambda_{max}} \xi_i = \frac{\lambda_{min} + \lambda_{max} - \lambda}{\lambda_{min} \lambda_{max}}, \end{aligned}$$

and it follows from (5.30) that

$$\begin{aligned} \frac{(y^T y)^2}{(y^T Q y)(y^T Q^{-1} y)} &= \frac{1}{\lambda \lambda_\phi} \geq \frac{\lambda_{min} \lambda_{max}}{\lambda (\lambda_{min} + \lambda_{max} - \lambda)} \\ &\geq \frac{\lambda_{min} \lambda_{max}}{\max_{\lambda \in [\lambda_{min}, \lambda_{max}]} \{\lambda (\lambda_{min} + \lambda_{max} - \lambda)\}} \\ &= \frac{4 \lambda_{min} \lambda_{max}}{(\lambda_{min} + \lambda_{max})^2}, \end{aligned}$$

which completes the proof. \square

Proposition 5.17 Let us consider the function $f(x) = \frac{1}{2}x^T Qx$, where Q is an $n \times n$ symmetric and positive definite matrix. Suppose that the steepest descent method applied to minimize the function f generates the iterates

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),$$

with α_k being determined by an exact line search

$$f(x_k - \alpha_k \nabla f(x_k)) = \min_{\alpha \geq 0} f(x_k - \alpha \nabla f(x_k)). \quad (5.31)$$

Then,

$$f(x_{k+1}) \leq \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 f(x_k), \quad \text{for all } k, \quad (5.32)$$

where $\lambda_{\min} = \lambda_{\min}(Q)$ and $\lambda_{\max} = \lambda_{\max}(Q)$.

Proof Since the gradient of the function f is

$$g_k := \nabla f(x_k) = Qx_k,$$

then (5.32) trivially holds when $g_k = 0_n$, as $x_{k+1} = x_k = 0_n$; accordingly, we shall assume that $g_k \neq 0_n$.

Let us start by calculating the stepsize. By (5.31),

$$\frac{d}{d\alpha} f(x_k - \alpha g_k) = -g_k^T Q(x_k - \alpha g_k) = -g_k^T g_k + \alpha g_k^T Q g_k.$$

Since the value of this derivative at α_k must be zero, one obtains

$$\alpha_k = \frac{g_k^T g_k}{g_k^T Q g_k}.$$

Then, as $Qx_k = g_k$, we get

$$\begin{aligned} f(x_{k+1}) &= \frac{1}{2}(x_k - \alpha_k g_k)^T Q(x_k - \alpha_k g_k) = \frac{1}{2}(x_k^T Q x_k - 2\alpha_k g_k^T Q x_k + \alpha_k^2 g_k^T Q g_k) \\ &= \frac{1}{2} \left(x_k^T Q x_k - \frac{(g_k^T g_k)^2}{g_k^T Q g_k} \right). \end{aligned}$$

Taking into account that

$$f(x_k) = \frac{1}{2}x_k^T Q x_k = \frac{1}{2}g_k^T Q^{-1} g_k,$$

we deduce, by applying Lemma 5.16,

$$\begin{aligned} f(x_{k+1}) &= \left(1 - \frac{(g_k^T g_k)^2}{(g_k^T Q g_k)(g_k^T Q^{-1} g_k)}\right) f(x_k) \\ &\leq \left(1 - \frac{4\lambda_{\max}\lambda_{\min}}{(\lambda_{\max} + \lambda_{\min})^2}\right) f(x_k) = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}\right)^2 f(x_k), \end{aligned}$$

which completes the proof. \square

It can be verified that the bounds (5.27) and (5.32) are accurate in the sense that their values are attained for certain functions and certain initial points. For instance, if $f(x) = \frac{1}{2} \sum_{i=1}^n \lambda_i x_i^2$ and $x_0 = (\lambda_{\min}^{-1}, 0, \dots, 0, \lambda_{\max}^{-1})^T$, with $0 < \lambda_{\min} = \lambda_1 \leq \dots \leq \lambda_n = \lambda_{\max}$. Any strictly convex quadratic form can be written in this way; see (3.3). Additional details for this example are given in [11, p. 68].

Finally, observe that every pair of consecutive directions of the steepest descent method with exact line search is orthogonal. Indeed, let $p_k = -\nabla f(x_k)$ and $p_{k+1} = -\nabla f(x_{k+1})$. By (5.31) and the chain rule, we have that

$$0 = \frac{df(x_k + \alpha p_k)}{d\alpha} \Big|_{\alpha=\alpha_k} = \nabla f(x_{k+1})^T p_k = -p_{k+1}^T p_k.$$

This explains the zigzagging behavior shown in Figs. 5.14 and 5.15.

5.2.3 Gradient Algorithms

A gradient algorithm applied to the function $f(x) = \frac{1}{2}x^T Qx$, where Q is an $n \times n$ symmetric and positive definite matrix produces iterates

$$x_{k+1} = x_k - \alpha_k B_k^{-1} \nabla f(x_k), \quad (5.33)$$

where B_k is an $n \times n$ symmetric and positive definite matrix. We show next that the iteration (5.33) is nothing else than the steepest descent method except for a transformation of coordinates.

We apply the change of coordinates $x = S_k y$, where

$$S_k = B_k^{-1/2}.$$

Recall that if ρ_1, \dots, ρ_n are the positive eigenvalues of B_k^{-1} and $\{u_1, \dots, u_n\}$ is an

orthonormal basis of associated eigenvectors, then

$$B_k^{-1/2} := \sum_{i=1}^n \rho_i^{1/2} u_i u_i^T$$

is a symmetric and positive definite matrix such that $B_k^{-1/2} B_k^{-1/2} = B_k^{-1}$; see (Section 1.2.2).

In the space of y -coordinates, our optimization problem is formulated as follows:

$$\begin{aligned} \text{Min } h_k(y) &:= f(S_k y) \\ \text{s.t. } y &\in \mathbb{R}^n. \end{aligned}$$

The steepest descent method applied to this problem produces iterates

$$y_{k+1} = y_k - \alpha_k \nabla h_k(y_k). \quad (5.34)$$

Multiplying by S_k , we get

$$S_k y_{k+1} = S_k y_k - \alpha_k S_k \nabla h_k(y_k),$$

and since $\nabla h_k(y_k) = S_k \nabla f(x_k)$ and $S_k^2 = B_k^{-1}$,

$$x_{k+1} = x_k - \alpha_k B_k^{-1} \nabla f(x_k).$$

Therefore, we have shown that the gradient method (5.33) is nothing else than the steepest descent algorithm (5.34) in the space of y -variables. This allows us to apply the results obtained in Proposition 5.17 and assert that

$$\frac{\|y_{k+1}\|}{\|y_k\|} \leq \max\{|1 - \alpha_k \lambda_{k,min}|, |1 - \alpha_k \lambda_{k,max}|\}, \quad (5.35)$$

and

$$\frac{f(x_{k+1})}{f(x_k)} = \frac{h_k(y_{k+1})}{h_k(y_k)} \leq \left(\frac{\lambda_{k,max} - \lambda_{k,min}}{\lambda_{k,max} + \lambda_{k,min}} \right)^2,$$

where $\lambda_{k,min}$ and $\lambda_{k,max}$ are the smallest and the largest eigenvalues of $\nabla^2 h_k(y)$, respectively, and

$$\nabla^2 h_k(y) = S_k \nabla^2 f(x) S_k = B_k^{-1/2} Q B_k^{-1/2}.$$

As $y_k = S_k^{-1} x_k = B_k^{1/2} x_k$, it follows from (5.35) that

$$\frac{x_{k+1}^T B_k x_{k+1}}{x_k^T B_k x_k} \leq \max\{(1 - \alpha_k \lambda_{k,min})^2, (1 - \alpha_k \lambda_{k,max})^2\}.$$

The stepsize minimizing this upper bound is

$$\frac{2}{\lambda_{k,\max} + \lambda_{k,\min}}. \quad (5.36)$$

An important observation is that, when B_k is a good approximation of $\nabla^2 f(x) = Q$, one has

$$\nabla^2 h_k(y) = B_k^{-1/2} Q B_k^{-1/2} \cong B_k^{-1/2} B_k B_k^{-1/2} = B_k^{-1/2} (B_k^{1/2} B_k^{1/2}) B_k^{-1/2} = I_n.$$

In such a case, we can expect that $\lambda_{k,\min} \cong 1 \cong \lambda_{k,\max}$. Moreover, the stepsize $\alpha_k = 1$ is close to be optimal (according to (5.36)).

5.3 Newton's Method

Newton's method generates successive iterations by the formula

$$x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k),$$

provided that the so-called *Newton's direction*

$$p_k^N := -(\nabla^2 f(x_k))^{-1} \nabla f(x_k), \quad (5.37)$$

is a well-defined descent direction; i.e., $\nabla f(x_k)^T p_k^N < 0$.

Newton's method is locally convergent; that is, it converges when x_0 is sufficiently close to a nonsingular local minimum \bar{x} . Its main drawback is that, when the number of variables n is large, the computational cost of $(\nabla^2 f(x_k))^{-1}$ is high.

In what follows we discuss the properties of the rate of convergence of Newton's method. Assuming that $f \in C^2$, if x is sufficiently close to a point \bar{x} such that $\nabla^2 f(\bar{x})$ is positive definite, then $\nabla^2 f(x)$ will also be positive definite. In this case, the *pure* Newton's method is clearly defined in this region and converges quadratically.

Theorem 5.18 (Convergence of Newton's method) *Assume that $\nabla^2 f$ is Lipschitz continuous on the closed ball $\bar{x} + \beta \mathbb{B}$, with \bar{x} being a point at which the second-order sufficient optimality conditions hold. Consider the iteration*

$$x_{k+1} = x_k + p_k^N,$$

where p_k^N is the direction in (5.37). Then, if the initial point x_0 is close enough to \bar{x} , the following statements are true:

- (i) The sequence $\{x_k\}$ generated by this algorithm quadratically converges to \bar{x} .
- (ii) The sequence $\{\|\nabla f_k\|\}$ also converges quadratically to zero.

Proof (i) From the definition of p_k^N and the first-order optimality condition $\nabla f_* := \nabla f(\bar{x}) = 0_n$, we have

$$\begin{aligned} x_k + p_k^N - \bar{x} &= x_k - \bar{x} - (\nabla^2 f_k)^{-1} \nabla f_k \\ &= (\nabla^2 f_k)^{-1} [(\nabla^2 f_k)(x_k - \bar{x}) - (\nabla f_k - \nabla f_*)]. \end{aligned} \quad (5.38)$$

Since (see (1.22))

$$\nabla f_k - \nabla f_* = \int_0^1 \nabla^2 f(\bar{x} + t(x_k - \bar{x}))(x_k - \bar{x}) dt,$$

we get

$$\begin{aligned} &\|(\nabla^2 f_k)(x_k - \bar{x}) - (\nabla f_k - \nabla f_*)\| \\ &= \left\| \int_0^1 [\nabla^2 f_k - \nabla^2 f(\bar{x} + t(x_k - \bar{x}))](x_k - \bar{x}) dt \right\| \\ &\leq \int_0^1 \|\nabla^2 f_k - \nabla^2 f(\bar{x} + t(x_k - \bar{x}))\| \|x_k - \bar{x}\| dt \\ &\leq \|x_k - \bar{x}\|^2 \int_0^1 L(1-t) dt = \frac{1}{2}L\|x_k - \bar{x}\|^2, \end{aligned} \quad (5.39)$$

if $x_k \in \bar{x} + \beta\mathbb{B}$, where L is the Lipschitz constant of $\nabla^2 f$ in this neighborhood of \bar{x} .

Thanks to the continuity of $(\nabla^2 f(x))^{-1}$, we can choose β small enough to guarantee that

$$\|(\nabla^2 f(x))^{-1}\| \leq 2 \|(\nabla^2 f(\bar{x}))^{-1}\| \quad (5.40)$$

for all $x \in \bar{x} + \beta\mathbb{B}$.

If $x_k \in \bar{x} + \beta\mathbb{B}$, by replacing this in (5.38) and using (5.39), we obtain

$$\begin{aligned} \|x_{k+1} - \bar{x}\| &= \|x_k + p_k^N - \bar{x}\| \leq L \|(\nabla^2 f(\bar{x}))^{-1}\| \|x_k - \bar{x}\|^2 \\ &= \tilde{L} \|x_k - \bar{x}\|^2, \end{aligned} \quad (5.41)$$

where $\tilde{L} := L \|(\nabla^2 f(\bar{x}))^{-1}\|$.

If we choose β sufficiently small to satisfy, in addition to (5.40), $\beta\tilde{L} < 1$, then

$$\|x_{k+1} - \bar{x}\| \leq \tilde{L} \|x_k - \bar{x}\| \|x_k - \bar{x}\| \leq \beta\tilde{L} \|x_k - \bar{x}\| \leq \|x_k - \bar{x}\| \leq \beta,$$

where we have taken into account that $x_k \in \bar{x} + \beta\mathbb{B}$.

Consequently, if $x_0 \in \bar{x} + \beta\mathbb{B}$, one concludes that $\{x_k\} \subset \bar{x} + \beta\mathbb{B}$, and in addition,

$$\|x_{k+1} - \bar{x}\| \leq \beta\tilde{L} \|x_k - \bar{x}\| \leq (\beta\tilde{L})^{k+1} \|x_0 - \bar{x}\|,$$

yielding the convergence $x_k \rightarrow \bar{x}$. The quadratic convergence follows from (5.41).

(ii) Taking into account that $x_{k+1} - x_k = p_k^N$ and $\nabla f_k + (\nabla^2 f_k)p_k^N = 0_n$, we obtain

$$\begin{aligned}\|\nabla f(x_{k+1})\| &= \|\nabla f(x_{k+1}) - \nabla f(x_k) - \nabla^2 f(x_k)p_k^N\| \\ &= \left\| \int_0^1 \nabla^2 f(x_k + tp_k^N)(x_{k+1} - x_k) dt - \nabla^2 f(x_k)p_k^N \right\| \\ &\leq \int_0^1 \|\nabla^2 f(x_k + tp_k^N) - \nabla^2 f(x_k)\| \|p_k^N\| dt \\ &\leq \frac{1}{2}L\|p_k^N\|^2 \leq \frac{1}{2}L\|\nabla^2 f(x_k)^{-1}\|^2\|\nabla f(x_k)\|^2 \\ &\leq 2L\|\nabla^2 f(\bar{x})^{-1}\|^2\|\nabla f(x_k)\|^2,\end{aligned}$$

where the last inequality follows from (5.40). Hence, we have proved that the sequence of gradient norms converge quadratically to zero. \square

The limitations of the pure Newton's method arise from the following facts:

- The convergence in the first iterations can be slow.
- The convergence to a local minimum can fail because:
 - The Hessian is singular (if $\nabla^2 f(x_k)$ is singular, then p_k^N is not defined).
 - The stepsize $\alpha_k = 1$ is too big and the quadratic approximation is less satisfactory if we are too far from x_k (see Fig. 5.16).

Our purpose now is to try to modify the pure Newton's method in order to force the global convergence while maintaining a good rate of local convergence. A simple possibility is to replace Newton's direction p_k^N by the steepest descent direction when p_k^N is not defined or is not a descent direction.

None of the variants of the pure Newton's method that have been established can ensure rapid convergence in the early iterations, but there are procedures that allow an effective use of second-order information, even when the Hessian is not positive definite. These schemes are based on changes in the diagonal elements of the Hessian. More precisely, the search direction p_k is obtained by solving the system

$$(\nabla^2 f(x_k) + \Delta_k)p_k = -\nabla f(x_k),$$

when Newton's direction p_k^N is not defined or is not a descent direction. Here, Δ_k is a diagonal matrix which is chosen in such a way that $\nabla^2 f(x_k) + \Delta_k$ is positive definite. We describe next one of the most well-known possibilities.

5.3.1 Trust Region Methods

Recall that the pure Newton's method is based on the minimization with respect to p of the quadratic approximation of f , around x_k , given by

$$f_k(p) := f(x_k) + \nabla f(x_k)^T p + \frac{1}{2} p^T \nabla^2 f(x_k) p.$$

We know that $f_k(p)$ is a good approximation of $f(x_k + p)$ whenever p is in a small neighborhood of 0_n . The problem is that the unrestricted minimization of $f_k(p)$ can lead to a new point $x_{k+1} = x_k + p_k$, with $p_k \in \operatorname{argmin}\{f_k(p) : p \in \mathbb{R}^n\}$, which is far from that neighborhood, as we have seen in Fig. 5.16.

It makes sense to consider a *restricted* step of Newton's method $x_{k+1} = x_k + p_k$ where p_k is obtained by minimizing $f_k(p)$ in a convenient neighborhood of 0_n called *trust region*; i.e.,

$$p_k \in \operatorname{argmin}\{f_k(p) : \|p\| \leq \gamma_k\},$$

with γ_k being a positive scalar (see Fig. 5.17).

By applying the KKT conditions (see Theorem 6.28, which is the nonconvex counterpart of Theorem 4.46), after formulating the restriction $\|p\| \leq \gamma_k$ as $\frac{1}{2}p^T I_n p \leq \frac{1}{2}\gamma_k^2$, it can be observed that the restricted Newton step, p_k , must be a solution of the system

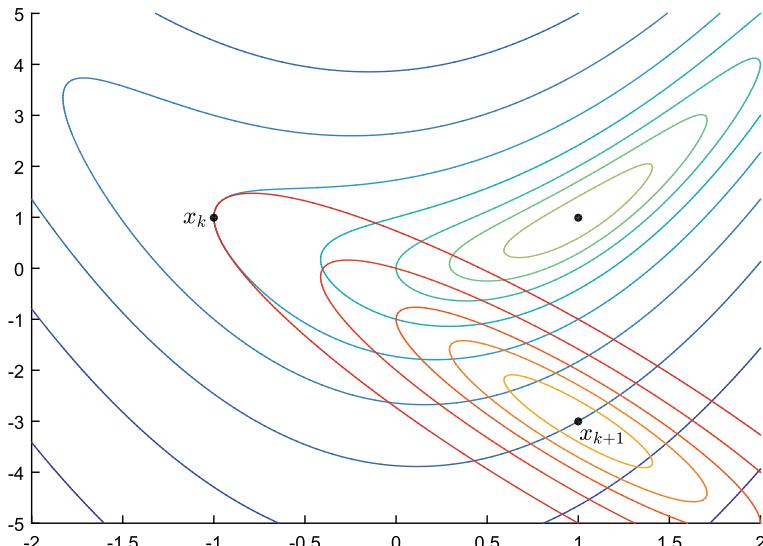


Fig. 5.16 Level curves of $f(x, y) = (y - x^2)^2 + (1 - x)^2$ (in blue-green) and of its quadratic approximation at x_k (in red-orange). The quadratic approximation is not satisfactory at x_{k+1} because the point is far from x_k . In fact, $f(x_{k+1}) > f(x_k)$

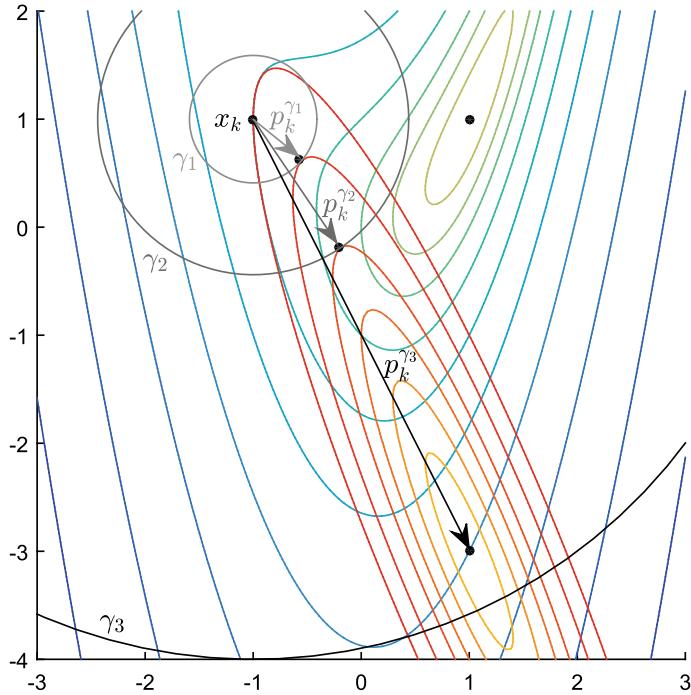


Fig. 5.17 Steps $p_k^{\gamma_1}, p_k^{\gamma_2}, p_k^{\gamma_3}$ for different radii $\gamma_1, \gamma_2, \gamma_3$ of the trust regions, showing the level lines of $f(x, y) = (y - x^2)^2 + (1 - x)^2$ (in blue-green) and of its quadratic approximation at x_k (in red-orange)

$$(\nabla^2 f(x_k) + \delta_k I_n) p_k = -\nabla f(x_k),$$

where δ_k is a nonnegative scalar (multiplier). Thus, it is evident that the present method of determining p_k corresponds to the strategy of using a diagonal correction of the Hessian matrix.

An important observation to make here is that even when $\nabla^2 f(x_k)$ is not positive definite, the restricted Newton step p_k improves the cost, provided that $\nabla f(x_k) \neq 0_n$ and γ_k is small enough. To check this claim, we note that for all p such that $\|p\| \leq \gamma_k$, we can write

$$f(x_k + p) = f_k(p) + o(\gamma_k^2),$$

and so

$$\begin{aligned} f(x_k + p_k) &= f_k(p_k) + o(\gamma_k^2) \\ &= f(x_k) + \min_{\|p\| \leq \gamma_k} \left\{ \nabla f(x_k)^T p + \frac{1}{2} p^T \nabla^2 f(x_k) p \right\} + o(\gamma_k^2). \end{aligned}$$

Hence, denoting by

$$\tilde{p}_k := -\frac{\nabla f(x_k)}{\|\nabla f(x_k)\|} \gamma_k,$$

one gets

$$\begin{aligned} f(x_{k+1}) &= f(x_k + p_k) \\ &\leq f(x_k) + \nabla f(x_k)^T \tilde{p}_k + \frac{1}{2} \tilde{p}_k^T \nabla^2 f(x_k) \tilde{p}_k + o(\gamma_k^2) \\ &= f(x_k) + \gamma_k \left(-\|\nabla f(x_k)\| + \frac{\gamma_k}{2\|\nabla f(x_k)\|^2} \nabla f(x_k)^T \nabla^2 f(x_k) \nabla f(x_k) + o(\gamma_k) \right). \end{aligned}$$

We observe that for γ_k sufficiently small, the term $-\|\nabla f(x_k)\|$ dominates the other two terms in the expression contained in the parenthesis, which shows that $f(x_{k+1}) < f(x_k)$.

The choice of the initial value of γ_k is crucial in this method. If it is chosen too large, perhaps many reductions of γ_k will be required until an improvement of the value of the objective function is achieved. If, however, the initial value of γ_k is too small, the rate of convergence can be very poor (see Fig. 5.17).

5.3.2 Least Squares Problems

This section deals with the optimization problem

$$\begin{aligned} P : \text{Min } f(x) &:= \frac{1}{2} \|g(x)\|^2 = \frac{1}{2} \sum_{i=1}^m g_i(x)^2 \\ \text{s.t. } x &\in \mathbb{R}^n, \end{aligned} \tag{5.42}$$

with $g = (g_1, \dots, g_m)^T : \mathbb{R}^n \rightarrow \mathbb{R}^m$, and $g_i \in \mathcal{C}^1$, $i = 1, 2, \dots, m$.

If our aim is to solve the equations system $g(x) = 0_m$, we can approach such a problem by minimizing the function $\frac{1}{2} \|g(x)\|^2$. It is evident that \bar{x} is a solution of the system if and only if \bar{x} is a global minimum of problem P and the optimal value is zero; i.e., $f(\bar{x}) = 0$.

Many other applications can be found in such diverse fields as curve fitting, neural networks, pattern classification, etc (see, for example, [11, pp. 93–97]).

We describe next the most commonly used method to solve problem (5.42), known as the *Gauss–Newton method*. Given a point x_k , the pure form of the Gauss–Newton method is based on linearizing the function g around the point x_k , that is, to consider the linear function

$$\begin{aligned}\varphi_k(x) &:= g(x_k) + \nabla g(x_k)^T(x - x_k) \\ &= g(x_k) + [\nabla g_1(x_k), \dots, \nabla g_m(x_k)]^T(x - x_k),\end{aligned}$$

and minimizing then the norm of this linear approximation. Thus, the iteration mechanism is the following:

$$\begin{aligned}x_{k+1} &= \operatorname{argmin} \left\{ \frac{1}{2} \|\varphi_k(x)\|^2 : x \in \mathbb{R}^n \right\} \\ &= \operatorname{argmin} \left\{ \frac{1}{2} \left\{ \begin{array}{l} \|g(x_k)\|^2 + 2g(x_k)^T \nabla g(x_k)^T (x - x_k) \\ + (x - x_k)^T \nabla g(x_k) \nabla g(x_k)^T (x - x_k) \end{array} \right\} : x \in \mathbb{R}^n \right\}.\end{aligned}\quad (5.43)$$

If the $n \times n$ matrix $\nabla g(x_k) \nabla g(x_k)^T$ is nonsingular, the objective function in (5.43) is quadratic and strictly convex; hence,

$$x_{k+1} = x_k - (\nabla g(x_k) \nabla g(x_k)^T)^{-1} \nabla g(x_k) g(x_k). \quad (5.44)$$

Note that if g is itself linear, one has $\|g(x)\|^2 = \|\varphi_k(x)\|^2$ and the method converges in a single iteration. Observe that the direction used in the iteration (5.44), i.e.,

$$p_k^{GN} := -(\nabla g(x_k) \nabla g(x_k)^T)^{-1} \nabla g(x_k) g(x_k),$$

is a descent direction since $\nabla g(x_k) g(x_k) = \sum_{i=1}^m g_i(x_k) \nabla g_i(x_k)$ is the gradient at x_k of the cost function $\frac{1}{2} \|g(x)\|^2$, and $(\nabla g(x_k) \nabla g(x_k)^T)^{-1}$ is a positive definite matrix (under the nonsingularity assumption).

To ensure that a descent occurs in the case that the matrix $\nabla g(x_k) \nabla g(x_k)^T$ is singular (and also to enhance the convergence when this matrix is close to be singular), it is usually performed an iteration of the type

$$x_{k+1} = x_k - \alpha_k (\nabla g(x_k) \nabla g(x_k)^T + \Delta_k)^{-1} \nabla g(x_k) g(x_k),$$

where α_k is chosen by means of any stepsize determination rule and Δ_k is a diagonal matrix for which

$$\nabla g(x_k) \nabla g(x_k)^T + \Delta_k$$

is positive definite. In the well-known *Levenberg–Marquardt method*, Δ_k is a positive multiple of I_n .

The Gauss–Newton method is closely related to Newton's method. In fact, the Hessian of the objective function is

$$\nabla g(x_k) \nabla g(x_k)^T + \sum_{i=1}^m g_i(x_k) \nabla^2 g_i(x_k),$$

so (5.44) is equivalent to one iteration of Newton's method but omitting the second-order term

$$\sum_{i=1}^m g_i(x_k) \nabla^2 g_i(x_k). \quad (5.45)$$

Thus, the Gauss–Newton method does not need the calculation of this term, and the price to be paid is a possible deterioration in the rate of convergence. Therefore, if the term (5.45) is relatively small, close to a minimum, the rate of convergence of the Gauss–Newton method is quite satisfactory. This will be particularly true in the case where g is practically linear, and also when the values of $g_i(x)$, $i = 1, \dots, m$, are small near the solution.

In the case where $m = n$, and we try to solve the system $g(x) = 0_n$, the omitted term (5.45) is zero at the solution. In this case, assuming that $\nabla g(x_k)$ is invertible, we have

$$(\nabla g(x_k) \nabla g(x_k)^T)^{-1} \nabla g(x_k) g(x_k) = (\nabla g(x_k)^T)^{-1} g(x_k),$$

and the pure form of the Gauss–Newton method (5.44) iterates as follows:

$$x_{k+1} = x_k - (\nabla g(x_k)^T)^{-1} g(x_k),$$

which is nothing else but Newton's method for solving $g(x) = 0_n$, as it is described in the following section.

5.4 Newton's Method for Solving Equations

With the aim of obtaining candidates for being local minimizers of a differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, we look for vectors \bar{x} satisfying $\nabla f(\bar{x}) = 0_n$. Observe that this equation is nonlinear if f is nonquadratic. In this section, we present Newton's method for approximately solving such nonlinear equation. The method is applied in Section 6.5, where the sequential quadratic methodology is studied.

Consider the differentiable function $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$. Starting from an initial point x_0 , the pure *Newton's method algorithm for solving the equation $g(x) = 0_n$* determines the iterates $x_1, x_2, \dots, x_k, x_{k+1}, \dots$ as follows. If $g(x_k) \neq 0_n$, we replace g by its first-order linear approximation

$$g(x) \approx g(x_k) + \nabla g(x_k)^T (x - x_k), \quad \text{for } x \text{ close to } x_k,$$

and choose x_{k+1} as a solution of the the system of linear equations

$$g(x_k) + \nabla g(x_k)^T (x - x_k) = 0_n,$$

or equivalently,

$$\nabla g(x_k)^T x = \nabla g(x_k)^T x_k - g(x_k).$$

Assuming that $\nabla g(x_k)$ is invertible, one gets the Newton step

$$x_{k+1} = x_k - (\nabla g(x_k)^T)^{-1} g(x_k). \quad (5.46)$$

Theorem 5.19 (Convergence of Newton's method for equations) *Let $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a $C^2(\mathbb{R}^n)$ function and \bar{x} be a zero of g , i.e., $g(\bar{x}) = 0_n$, such that $\det(\nabla g(\bar{x})) \neq 0$. Then, the Newton iteration (5.46) converges quadratically to \bar{x} provided that x_0 is close enough to \bar{x} .*

Proof The proof is quite similar to the proof of Theorem 5.18. By the continuity assumption (and the continuity of the norm), there exists $\beta > 0$, sufficiently small to satisfy $\det(\nabla g(x)) \neq 0$ and

$$\|(\nabla g(x)^T)^{-1}\| \leq 2 \|(\nabla g(\bar{x})^T)^{-1}\|, \quad \forall x \in \bar{x} + \beta \mathbb{B}. \quad (5.47)$$

Then, if $\|x_k - \bar{x}\| < \beta$, we derive from (5.46)

$$\begin{aligned} x_{k+1} - \bar{x} &= x_k - \bar{x} - (\nabla g(x_k)^T)^{-1} g(x_k) \\ &= (\nabla g(x_k)^T)^{-1} (\nabla g(x_k)^T (x_k - \bar{x}) - (g(x_k) - g(\bar{x}))). \end{aligned} \quad (5.48)$$

Since

$$g(x_k) - g(\bar{x}) = \int_0^1 \nabla g(\bar{x} + t(x_k - \bar{x}))^T (x_k - \bar{x}) dt,$$

one has

$$\begin{aligned} &\|\nabla g(x_k)^T (x_k - \bar{x}) - (g(x_k) - g(\bar{x}))\| \\ &= \left\| \int_0^1 (\nabla g(x_k) - \nabla g(\bar{x} + t(x_k - \bar{x})))^T (x_k - \bar{x}) dt \right\| \\ &\leq \int_0^1 \|\nabla g(x_k) - \nabla g(\bar{x} + t(x_k - \bar{x}))\| \|x_k - \bar{x}\| dt. \end{aligned} \quad (5.49)$$

Applying Proposition 5.7 to the components of g , say g_i , $i = 1, \dots, m$, with $U := \bar{x} + \beta \text{int } \mathbb{B}$, and thanks to the fact that the Hessians $\nabla^2 g_i$, $i = 1, \dots, m$, are continuous and, therefore, bounded on the ball $\bar{x} + \beta \mathbb{B}$, we conclude that ∇g is Lipschitz continuous on U . Thus, there exists a Lipschitz constant $L > 0$ such that

$$\|\nabla g(x) - \nabla g(y)\| \leq L \|x - y\|, \quad \forall x, y \in U.$$

Now, from (5.49), we obtain

$$\begin{aligned}\|\nabla g(x_k)^T(x_k - \bar{x}) - (g(x_k) - g(\bar{x}))\| &\leq \|x_k - \bar{x}\|^2 \int_0^1 L(1-t)dt \\ &= \frac{1}{2}L\|x_k - \bar{x}\|^2.\end{aligned}\quad (5.50)$$

Using (5.47), (5.48), and (5.50), one gets

$$\|x_{k+1} - \bar{x}\| \leq L \left\| (\nabla g(\bar{x})^T)^{-1} \right\| \|x_k - \bar{x}\|^2. \quad (5.51)$$

Like in Theorem 5.18, if β is taken small enough to additionally satisfy

$$\beta L \left\| (\nabla g(\bar{x})^T)^{-1} \right\| < 1,$$

then we conclude that the whole sequence of iterates $\{x_k\}$ is contained in $\bar{x} + \beta\mathbb{B}$ and converges to \bar{x} . The quadratic convergence comes from (5.51). \square

Example 5.20 Consider the system of equations

$$\begin{aligned}x + 6y - 3xy &= 0, \\ x^2y^2 + 4xy^2 - 10xy + 5y^2 - 20y + 15 &= 0.\end{aligned}\quad (5.52)$$

For any starting point $(x_0, y_0)^T \in \mathbb{R}^2$, the iterates of the pure Newton's method for the system (5.52) are given by

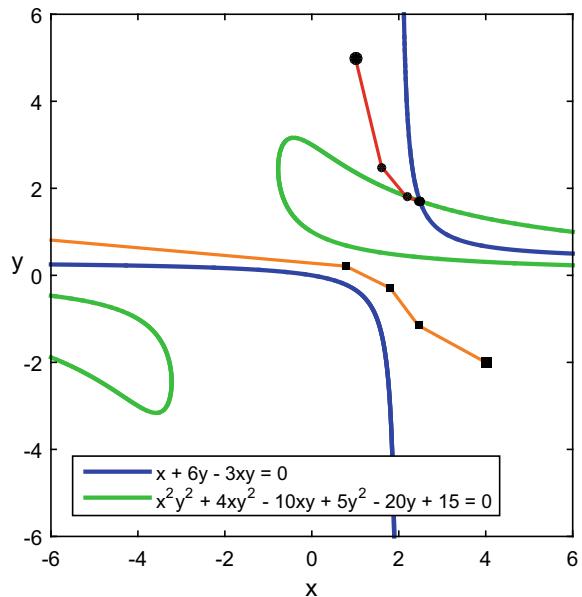
$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} - \begin{bmatrix} 1 - 3y_k & 2x_k y_k^2 + 4y_k^2 - 10y_k \\ 6 - 3x_k & 2x_k^2 y_k + 8x_k y_k - 10x_k + 10y_k - 20 \end{bmatrix}^{-1} \begin{pmatrix} x_k + 6y_k - 3x_k y_k \\ x_k^2 y_k^2 + 4x_k y_k^2 - 10x_k y_k + 5y_k^2 - 20y_k + 15 \end{pmatrix}.$$

In Fig. 5.18, we show Newton's sequence for two different starting points. The sequence generated by $(x_0, y_0) = (1, 5)$ (in red) converges to the solution of the equation $(\bar{x}, \bar{y}) = (2.4848866, 1.7082252)$. Since $\det(\nabla g(\bar{x})) = -99.3235716$, the convergence rate is guaranteed to be quadratic, by Theorem 5.19. On the other hand, for the starting point $(\tilde{x}_0, \tilde{y}_0) = (4, -2)$ (in orange), the sequence diverges. This shows the relevance of the proximity assumption to the solution in Theorem 5.19.

5.5 Conjugate Direction Methods

The purpose of this family of methods is to improve the rate of convergence of the steepest descent method, without incurring the computational burden of Newton's method. Originally, they were developed to solve the quadratic problem

Fig. 5.18 A convergent (red) and a divergent (orange) sequence generated by Newton's method for two different starting points



$$\begin{aligned} \text{Min } f(x) &= \frac{1}{2}x^T Qx - c^T x \\ \text{s.t. } x &\in \mathbb{R}^n, \end{aligned} \quad (5.53)$$

where Q is a symmetric and positive definite matrix, or to solve the linear system

$$Qx = c.$$

Conjugate direction methods solve these problems in a maximum of n iterations. They can also be applied to optimization problems in a neighborhood of a nonsingular local minimum (see [11, p. 118]).

Definition 5.21 Let Q be an $n \times n$ symmetric and positive definite matrix. The nonzero vectors p_0, p_1, \dots, p_k define a set of Q -conjugate directions if

$$p_i^T Q p_j = 0, \quad \forall i, j \in \{1, 2, \dots, k\}, i \neq j.$$

Lemma 5.22 If p_0, p_1, \dots, p_k are Q -conjugate, then they are linearly independent.

Proof Otherwise, we can suppose without loss of generality that

$$p_0 = t_1 p_1 + \dots + t_k p_k.$$

Then,

$$p_0^T Q p_0 = \sum_{i=1}^k t_i p_i^T Q p_0 = 0,$$

because $p_i^T Q p_0 = 0$, $i = 1, \dots, k$. Since $p_0 \neq 0_n$ and Q is positive definite, we get a contradiction. \square

Starting from a (maximal) set of n Q -conjugate directions, p_0, p_1, \dots, p_{n-1} , the *method of conjugate directions*, aimed at solving the problem (5.53), works through the iterative scheme given by

$$x_{k+1} = x_k + \alpha_k p_k, \quad k = 0, 1, \dots, n-1,$$

where x_0 is an arbitrary initial point and α_k is obtained by an exact line search on the whole straight line determined by p_k and passing through x_k , i.e.,

$$f(x_k + \alpha_k p_k) = \min\{f(x_k + \alpha p_k) : \alpha \in \mathbb{R}\}. \quad (5.54)$$

Proposition 5.23 *For each k , we have*

$$x_{k+1} = \operatorname{argmin}\{f(x) : x \in M_k\},$$

where

$$M_k := x_0 + \operatorname{span}\{p_0, p_1, \dots, p_k\}.$$

In particular, x_n minimizes f over the whole space \mathbb{R}^n , since $M_{n-1} = \mathbb{R}^n$.

Proof The definition of α_k in (5.54) yields

$$\frac{df(x_k + \alpha p_k)}{d\alpha} \Big|_{\alpha=\alpha_k} = \nabla f(x_{k+1})^T p_k = 0,$$

and, for $i = 0, 1, \dots, k-1$, we have

$$\begin{aligned} \nabla f(x_{k+1})^T p_i &= (Qx_{k+1} - c)^T p_i = \left(x_{i+1} + \sum_{j=i+1}^k \alpha_j p_j \right)^T Q p_i - c^T p_i \\ &= x_{i+1}^T Q p_i - c^T p_i = (Qx_{i+1} - c)^T p_i = \nabla f(x_{i+1})^T p_i, \end{aligned}$$

where the Q -conjugacy of p_i and p_j , $j = i+1, \dots, k$, has been taken into account. Combining these equalities, one gets

$$\nabla f(x_{k+1})^T p_i = 0, \quad i = 0, 1, \dots, k. \quad (5.55)$$

In this way,

$$\frac{\partial f(x_0 + \gamma_0 p_0 + \dots + \gamma_k p_k)}{\partial \gamma_i} \Big|_{\gamma_j=\alpha_j, j=0,1,\dots,k} = 0, \quad i = 0, \dots, k,$$

entailing that the strictly convex function

$$\varphi(\gamma_0, \dots, \gamma_k) := f(x_0 + \gamma_0 p_0 + \dots + \gamma_k p_k)$$

satisfies

$$\nabla \varphi(\alpha_0, \dots, \alpha_k) = 0_{k+1},$$

leading us to the aimed conclusion. \square

Given a set of linearly independent vectors $\{v_0, v_1, \dots, v_k\}$, now we set the task of building a set of Q -conjugate directions $\{p_0, p_1, \dots, p_k\}$ such that

$$\text{span}\{p_0, p_1, \dots, p_k\} = \text{span}\{v_0, v_1, \dots, v_k\}.$$

To this aim, we use the following variant of the Gram–Schmidt method. The method is recursive and works as follows:

(1) It starts with

$$p_0 = v_0.$$

(2) Assume that, for some $i < k$, the Q -conjugate directions p_0, p_1, \dots, p_i satisfying

$$\text{span}\{p_0, p_1, \dots, p_i\} = \text{span}\{v_0, v_1, \dots, v_i\}, \quad (5.56)$$

have already been built.

(3) Introduce a new direction

$$p_{i+1} := v_{i+1} + \sum_{m=0}^i \lambda_{i+1,m} p_m, \quad (5.57)$$

choosing the coefficients $\lambda_{i+1,m}$, $m = 0, 1, \dots, i$, in such a way that p_{i+1} is Q -conjugate to p_0, p_1, \dots, p_i .

Since p_{i+1} is required to be Q -conjugate to the previous directions p_0, p_1, \dots, p_i , for each $j = 0, 1, \dots, i$, it must be

$$0 = p_{i+1}^T Q p_j = v_{i+1}^T Q p_j + \sum_{m=0}^i \lambda_{i+1,m} p_m^T Q p_j = v_{i+1}^T Q p_j + \lambda_{i+1,j} p_j^T Q p_j,$$

yielding

$$\lambda_{i+1,j} = -\frac{v_{i+1}^T Q p_j}{p_j^T Q p_j}, \quad j = 0, 1, \dots, i. \quad (5.58)$$

Let us observe that the denominator $p_j^T Q p_j$ is positive because the directions p_0, p_1, \dots, p_i are (due to the induction hypothesis) Q -conjugate and thus nonzero.

Note also that $p_{i+1} \neq 0_n$. Otherwise, if $p_{i+1} = 0_n$, we would have by (5.57) and (5.56) that

$$v_{i+1} \in \text{span}\{p_0, p_1, \dots, p_i\} = \text{span}\{v_0, v_1, \dots, v_i\},$$

which contradicts the linear independence of vectors v_0, v_1, \dots, v_k .

Finally, by (5.57),

$$v_{i+1} \in \text{span}\{p_0, p_1, \dots, p_i, p_{i+1}\},$$

meanwhile

$$\begin{aligned} p_{i+1} &\in \text{span}\{p_0, p_1, \dots, p_i\} + \text{span}\{v_{i+1}\} \\ &= \text{span}\{v_0, v_1, \dots, v_i\} + \text{span}\{v_{i+1}\} \\ &= \text{span}\{v_0, v_1, \dots, v_i, v_{i+1}\}. \end{aligned}$$

Hence, (5.56) holds when i increases to $i + 1$.

It is also worth examining the case in which the vectors v_0, v_1, \dots, v_i are linearly independent, but the vector v_{i+1} is linearly dependent to them. In this case, the above procedure (5.57) and the formulas (5.58) remain valid, but the new vector p_{i+1} will be null. Indeed, from (5.56) and (5.57), one has

$$p_{i+1} \in \text{span}\{v_0, v_1, \dots, v_i, v_{i+1}\} = \text{span}\{v_0, v_1, \dots, v_i\},$$

and

$$p_{i+1} = \sum_{m=0}^i \gamma_m p_m. \quad (5.59)$$

Multiplying (5.59) by $p_j^T Q$, $j = 0, 1, \dots, i$, it turns out that $\gamma_m = 0$, $m = 0, 1, \dots, i$, and $p_{i+1} = 0_n$.

We can use this property to build a set of Q -conjugate directions that generate the same subspace that the vectors v_0, v_1, \dots, v_k , which a priori need not be linearly independent. Whenever after using (5.57) and (5.58), the new generated direction p_{i+1} is null, it will be discarded and we shall introduce v_{i+2} , and so on.

5.5.1 Conjugate Gradient Method

This method consists of applying the variant of the Gram–Schmidt method described above to the vectors

$$v_k = -g_k \equiv -\nabla f(x_k) = -(Qx_k - c), \quad k = 0, 1, \dots, n - 1.$$

Thus, the conjugate gradient method generates the iterates

$$x_{k+1} = x_k + \alpha_k p_k,$$

where α_k minimizes f over the straight line $\{x_k + \alpha p_k : \alpha \in \mathbb{R}\}$, and p_k is obtained by applying (5.57) to $-g_k$ and the previously generated directions p_0, p_1, \dots, p_{k-1} with coefficients provided by (5.58):

$$p_k = -g_k + \sum_{j=0}^{k-1} \frac{g_k^T Q p_j}{p_j^T Q p_j} p_j. \quad (5.60)$$

Note that $p_0 = -g_0$ and the method ends when it reaches a point x_k such that $g_k = 0_n$. Obviously, the method also stops when $p_k = 0_n$, but we will see that this can only happen when $g_k = 0_n$.

The key property of the conjugate gradient method is that the formula (5.60) can be considerably simplified. In particular, all but one of the coefficients in (5.60) vanish. This is a consequence of (5.55), which states that the gradient g_k is orthogonal to p_0, p_1, \dots, p_{k-1} . In fact, we have the following result.

Proposition 5.24 *The search directions used by the conjugate gradient method are*

$$\begin{aligned} p_0 &= -g_0, \\ p_k &= -g_k + \beta_k p_{k-1}, \quad k = 1, 2, \dots, n-1, \end{aligned}$$

with

$$\beta_k := \frac{g_k^T g_k}{g_{k-1}^T g_{k-1}}. \quad (5.61)$$

Furthermore, the method reaches an optimal solution after at most n steps.

Proof We will use induction to prove that the gradients g_k that are generated before termination are linearly independent. The result is obvious if $k = 0$ and $g_0 \neq 0_n$. Suppose then that the method has not finished after k -step stages, and that g_0, g_1, \dots, g_{k-1} are linearly independent. Then, since this is a method of conjugate directions,

$$\text{span}\{p_0, p_1, \dots, p_{k-1}\} = \text{span}\{g_0, g_1, \dots, g_{k-1}\}.$$

Two possibilities arise:

- (i) $g_k = 0_n$, in which case the method stops;
- (ii) $g_k \neq 0_n$, in which case, by (5.55),

$$g_k \text{ is orthogonal to } \text{span}\{p_0, p_1, \dots, p_{k-1}\} = \text{span}\{g_0, g_1, \dots, g_{k-1}\}. \quad (5.62)$$

and this entails that g_k is linearly independent to $\{g_0, g_1, \dots, g_{k-1}\}$ (otherwise, g_k would be orthogonal to itself, i.e., $g_k = 0_n$).

Since at most n linearly independent gradients can be generated, it follows that the gradient will be 0_n after n iterations, and the method terminates obtaining the (global) minimum of f (see Fig. 5.19).

Let us see now that (5.60) adopts the indicated simple form. If j is such that $g_j \neq 0_n$, we have

$$g_{j+1} - g_j = Q(x_{j+1} - x_j) = \alpha_j Q p_j. \quad (5.63)$$

Observe that $\alpha_j \neq 0$, since otherwise $g_{j+1} = g_j$, implying (in virtue of (5.62)) that $g_{j+1} = g_j = 0_n$ (which is discarded by assumption). Thus,

$$g_i^T Q p_j = \frac{1}{\alpha_j} g_i^T (g_{j+1} - g_j) = \begin{cases} 0, & \text{if } j = 0, 1, \dots, i-2, \\ \frac{1}{\alpha_{i-1}} g_i^T g_i, & \text{if } j = i-1, \end{cases}$$

and also

$$p_j^T Q p_j = \frac{1}{\alpha_j} p_j^T (g_{j+1} - g_j).$$

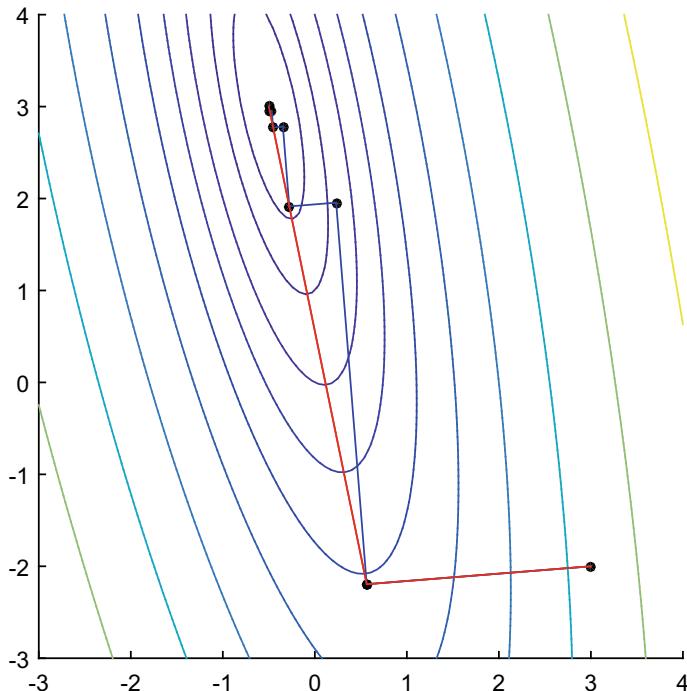


Fig. 5.19 Comparison of steepest descent method (blue) with the conjugate gradient method (red), which converges in a maximum of $n = 2$ iterations

Replacing this in (5.60) gives rise to

$$p_k = -g_k + \beta_k p_{k-1}, \quad (5.64)$$

with

$$\beta_k = \frac{\frac{1}{\alpha_{k-1}} g_k^T g_k}{\frac{1}{\alpha_{k-1}} p_{k-1}^T (g_k - g_{k-1})} = \frac{g_k^T g_k}{p_{k-1}^T (g_k - g_{k-1})}. \quad (5.65)$$

We deduce from (5.64) that

$$p_{k-1} = -g_{k-1} + \beta_{k-1} p_{k-2}.$$

Using this equality, the orthogonality of g_k and g_{k-1} , and of p_{k-2} and $g_k - g_{k-1}$ (by (5.62) and (5.63)), we see that the denominator of (5.65) collapses to $g_{k-1}^T g_{k-1}$, as we wanted to show. \square

Although the conjugate gradient method was designed to solve the quadratic problem (5.53), it can be adapted to minimize nonquadratic functions. Let us give some insights on how the scheme can be extended for solving the nonquadratic optimization problem

$$\begin{aligned} \text{Min } & f(x), \\ \text{s.t. } & x \in \mathbb{R}^n. \end{aligned}$$

The iteration is also given by

$$x_{k+1} = x_k + \alpha_k p_k,$$

with $p_0 = -\nabla f(x_0)$, and α_k is again obtained by an exact line search

$$f(x_k + \alpha_k p_k) = \min\{f(x_k + \alpha p_k), \alpha \in \mathbb{R}\}, \quad (5.66)$$

and the new direction is defined by

$$p_k := -\nabla f(x_k) + \beta_k p_{k-1}, \quad (5.67)$$

where the most common way of calculating β_k is through the formula

$$\beta_k = \frac{\nabla f(x_k)^T (\nabla f(x_k) - \nabla f(x_{k-1}))}{\nabla f(x_{k-1})^T \nabla f(x_{k-1})}. \quad (5.68)$$

At first sight, this formula is slightly different from (5.61). Nonetheless, observe that, in the quadratic case, the orthogonality of g_k and g_{k-1} allows to write formula (5.61) as

$$\beta_k := \frac{g_k^T(g_k - g_{k-1})}{g_{k-1}^T g_{k-1}}, \quad (5.69)$$

which is similar to (5.68). Whereas (5.61) and (5.69) are equivalent in the quadratic case, in the nonquadratic case there are differences between the methods defined by both formulas.

Finally, let us remark that the vector p_k given by (5.67) is a descent direction:

$$\nabla f(x_k)^T p_k = -\|\nabla f(x_k)\|^2 + \beta_k \nabla f(x_k)^T p_{k-1} = -\|\nabla f(x_k)\|^2,$$

where the first equality comes from the definition (5.67) and the second one is a consequence of (5.66) for the iteration $k - 1$.

The conjugate gradient method is often used in problems where the number of variables n is large. Frequently, the method suddenly starts to generate inefficient search directions. For this reason, it is important to restart the method after every n iterations by setting $\beta_k = 0$. In this way, we operate in cycles of steps using conjugate directions, with the first iteration in the cycle performed by the steepest descent method. This is important because the finite termination property for quadratic functions only holds when the initial direction is equal to the negative gradient. Thus, if the iterations generated by the algorithm converge to a neighborhood of a solution where the function is strictly convex quadratic, at some step the algorithm will be restarted, and finite termination will occur.

5.5.2 Quasi-Newton Methods

They are gradient methods that iterate as follows:

$$x_{k+1} = x_k + \alpha_k p_k,$$

with

$$p_k := -D_k \nabla f(x_k) \text{ and } \alpha_k > 0, \quad (5.70)$$

where D_k is a symmetric and positive definite matrix, which is updated at each iteration in such a way that p_k is progressively approaching Newton's direction p_k^N , see (5.37), meanwhile D_k approaches $(\nabla^2 f(x_k))^{-1}$.

Typically, their convergence is fast, and they avoid the calculation of the second derivatives involved in the application of Newton's method. These methods require the storage of the matrix D_k and of the other elements involved in obtaining D_{k+1} from D_k .

A fundamental idea behind the quasi-Newton methods is that every two consecutive points, x_k and x_{k+1} , together with their gradients, $\nabla f(x_k)$ and $\nabla f(x_{k+1})$, provide information about the curvature of f , through the approximate relationship

$$q_k \approx \nabla^2 f(x_{k+1})s_k, \quad (5.71)$$

where

$$s_k := x_{k+1} - x_k \quad \text{and} \quad q_k := \nabla f(x_{k+1}) - \nabla f(x_k).$$

Observe that if f is a quadratic function, $\nabla^2 f$ is constant and (5.71) becomes an equality.

As we mentioned before, when we apply a quasi-Newton method, $(\nabla^2 f(x_k))^{-1}$ is approximated by D_k , a matrix which is required to be symmetric and positive definite, for every k , and to satisfy the following relationship which plays the role of (5.71):

$$s_k = D_{k+1}q_k. \quad (5.72)$$

The matrix D_{k+1} is obtained from D_k , s_k and q_k through the relation

$$D_{k+1} = D_k + E_k, \quad (5.73)$$

where the *updating matrix* E_k is a *low-rank* symmetric and positive definite matrix. A possibility consists of taking

$$E_k = \gamma aa^T + \beta bb^T, \quad \text{with } a, b \in \mathbb{R}^n, \quad (5.74)$$

in which case (5.72) and (5.73) imply

$$s_k = D_{k+1}q_k = D_kq_k + E_kq_k = D_kq_k + \gamma aa^T q_k + \beta bb^T q_k. \quad (5.75)$$

Condition (5.75) holds if we choose

$$a = s_k \quad \text{and} \quad b = D_kq_k, \quad (5.76)$$

together with

$$\gamma = \frac{1}{s_k^T q_k} \quad \text{and} \quad \beta = -\frac{1}{q_k^T D_k q_k}. \quad (5.77)$$

Indeed, replacing (5.76) and (5.77) in (5.75) one gets

$$\begin{aligned} D_{k+1}q_k &= D_kq_k + \gamma aa^T q_k + \beta bb^T q_k \\ &= D_kq_k + \frac{1}{s_k^T q_k} s_k (s_k^T q_k) - \frac{1}{q_k^T D_k q_k} D_k q_k (q_k^T D_k q_k) \\ &= D_kq_k + s_k - D_k q_k \\ &= s_k. \end{aligned}$$

With this choice, the update (5.74)

$$E_k = \gamma aa^T + \beta bb^T = \frac{s_k s_k^T}{s_k^T q_k} - \frac{D_k q_k q_k^T D_k}{q_k^T D_k q_k}. \quad (5.78)$$

This is the well-known *Davidon–Fletcher–Powell (DFP) method*, which was introduced by Davidon in 1959 and later developed by Fletcher and Powell, and is historically the first quasi-Newton method.

If $\nabla f(x_k) \neq 0_n$ and D_k is definite positive, the vector p_k given by (5.70) is a descent direction. Moreover, if $\alpha_k > 0$ is chosen in such a way that

$$\nabla f(x_k)^T p_k < \nabla f(x_{k+1})^T p_k, \quad (5.79)$$

then $q_k \neq 0_n$, entailing $q_k^T D_k q_k > 0$. In particular, if α_k is obtained by an exact line search over the half-line $\{x_k + \alpha p_k : \alpha > 0\}$, we shall have $\nabla f(x_{k+1})^T p_k = 0$ and (5.79) trivially holds. Additionally, thanks to (5.79) and the fact that $\alpha_k > 0$,

$$s_k^T q_k = \alpha_k p_k^T (\nabla f(x_{k+1}) - \nabla f(x_k)) > 0.$$

Thus, under (5.79), we conclude that the denominators in (5.77) are both positive, and from (5.78), it is possible to deduce (see Proposition 5.25) that if D_k is symmetric and positive definite, D_{k+1} enjoys the same properties.

Let us explore next the possibility of considering in (5.74) an additional term and take

$$E_k = \gamma aa^T + \beta bb^T + \delta(ab^T + ba^T) = (a, b) \begin{pmatrix} \gamma & \delta \\ \delta & \beta \end{pmatrix} \begin{pmatrix} a^T \\ b^T \end{pmatrix}, \quad (5.80)$$

with $a = s_k$ and $b = D_k q_k$, like in (5.76), and parameters γ , β , and δ to be fixed later on.

Obviously, E_k is symmetric and D_{k+1} will inherit this property from D_k . We see that the rank of E_k is not greater than two since for all $z \in \mathbb{R}^n$,

$$E_k z = (\gamma a^T z + \delta b^T z)a + (\delta a^T z + \beta b^T z)b \in \text{span}\{a, b\},$$

and E_k is of rank 2 if and only if $\gamma a + \delta b$ and $\delta a + \beta b$ are linearly independent (entailing that a and b are also linearly independent).

Next, we proceed to choose values for γ , β , and δ in (5.80) such that (5.72) and (5.73) hold if we take again $a = s_k$ and $b = D_k q_k$. With such an assignment, it must be:

$$\begin{aligned} a &= s_k = D_{k+1} q_k = D_k q_k + (a, b) \begin{pmatrix} \gamma & \delta \\ \delta & \beta \end{pmatrix} \begin{pmatrix} a^T q_k \\ b^T q_k \end{pmatrix} \\ &= b + (\gamma a^T q_k + \delta b^T q_k)a + (\delta a^T q_k + \beta b^T q_k)b, \end{aligned}$$

or, equivalently,

$$0_n = (\gamma a^T q_k + \delta b^T q_k - 1)a + (\delta a^T q_k + \beta b^T q_k + 1)b. \quad (5.81)$$

The fulfillment of (5.81) is guaranteed if we choose γ , β , and δ satisfying

$$\begin{aligned} \gamma a^T q_k + \delta b^T q_k &= 1, \\ \delta a^T q_k + \beta b^T q_k &= -1, \end{aligned}$$

or, equivalently, if

$$\begin{aligned} \gamma s_k^T q_k + \delta q_k^T D_k q_k &= 1, \\ \delta s_k^T q_k + \beta q_k^T D_k q_k &= -1. \end{aligned}$$

We have a system of two equations with three unknowns, and we can choose the value of a parameter arbitrarily. In particular, taking $\beta = 0$, we get from the second equation

$$\delta = -\frac{1}{s_k^T q_k},$$

and from the first one

$$\gamma = \frac{s_k^T q_k + q_k^T D_k q_k}{(s_k^T q_k)^2}.$$

Replacing this in (5.80), we get

$$\begin{aligned} E_k &= \gamma aa^T + \beta bb^T + \delta(ab^T + ba^T) \\ &= \frac{s_k^T q_k + q_k^T D_k q_k}{(s_k^T q_k)^2} s_k s_k^T - \frac{1}{s_k^T q_k} (s_k q_k^T D_k + D_k q_k s_k^T) \\ &= \frac{s_k s_k^T}{s_k^T q_k} - \frac{D_k q_k q_k^T D_k}{q_k^T D_k q_k} + w_k w_k^T, \end{aligned} \quad (5.82)$$

where

$$w_k := (q_k^T D_k q_k)^{1/2} \left(\frac{s_k}{s_k^T q_k} - \frac{D_k q_k}{q_k^T D_k q_k} \right). \quad (5.83)$$

This expression differs from (5.80) in the term $w_k w_k^T$ and corresponds to the *Broyden–Fletcher–Goldfarb–Shanno (BFGS) method*, considered the best quasi-Newton method known to date (of general purpose).

We prove next that, under a weak condition, the matrices D_k generated with the updating matrix E_k given either by (5.78) or by (5.82) and (5.83) are positive definite. This ensures that the search direction p_k given by (5.70) is a descent direction.

Proposition 5.25 *If D_k is a symmetric and positive definite matrix and $\alpha_k > 0$ satisfies (5.79), then $D_{k+1} = D_k + E_k$ with E_k given by either (5.78) or (5.82) and (5.83) is also symmetric and positive definite.*

Proof We give the proof for the BFGS formula. We already established that, under the current assumptions, all the denominators in (5.78), (5.82), and (5.83) are nonzero (in fact, they are positive), and D_{k+1} is well-defined.

Now, for every $z \neq 0_n$, we have

$$z^T D_{k+1} z = z^T D_k z + \frac{(z^T s_k)^2}{s_k^T q_k} - \frac{(q_k^T D_k z)^2}{q_k^T D_k q_k} + (w_k^T z)^2. \quad (5.84)$$

Defining

$$u := D_k^{1/2} z \quad \text{and} \quad v := D_k^{1/2} q_k,$$

The equation (5.84) can be rewritten as

$$z^T D_{k+1} z = \frac{\|u\|^2 \|v\|^2 - (u^T v)^2}{\|v\|^2} + \frac{(z^T s_k)^2}{s_k^T q_k} + (w_k^T z)^2. \quad (5.85)$$

Since $s_k^T q_k$ is positive, from the Cauchy–Schwarz inequality we see that all the terms in the right-hand side of (5.85) are nonnegative. To prove that $z^T D_{k+1} z$ is, in fact, positive we show that

$$\|u\|^2 \|v\|^2 = (u^T v)^2 \quad \text{and} \quad z^T s_k = 0,$$

cannot hold simultaneously. Actually, $\|u\|^2 \|v\|^2 = (u^T v)^2$ entails $u = \lambda v$ or, equivalently,

$$z = \lambda q_k.$$

Since $z \neq 0_n$, it follows that $\lambda \neq 0$. Now, if $z^T s_k = 0$, it must be $q_k^T s_k = 0$, but this was already discarded. \square

It happens that (5.80) ensures that any quasi-Newton method that uses this type of correction is a method of conjugate directions when applied to strictly convex quadratic functions; see [33].

Proposition 5.26 *Let Q be a symmetric and positive definite matrix and consider the quadratic function*

$$f(x) = \frac{1}{2} x^T Q x - c^T x.$$

Suppose that we apply a quasi-Newton method to minimize f , with an updating matrix E_k which is either (5.78) or (5.82) and (5.83), and that we perform an exact line search at each iteration. Assume also that, starting from a nonoptimal point

x_0 and D_0 (symmetric and positive definite), the algorithm generates the iterates x_1, \dots, x_{n-1} , and that none of these points is optimal. Then, the generated directions $p_k := -D_k \nabla f(x_k)$, $k = 0, 1, \dots, n-1$, are Q -conjugate, i.e.,

$$p_i^T Q p_j = 0, \quad \forall i < j \quad \text{and} \quad D_n = Q^{-1}.$$

Proof We will prove it for the formula (5.78), i.e., for the DFP method. By induction over k , with $k = 0, 1, \dots, n-1$, we shall show that

$$p_i^T Q p_j = 0, \quad 0 \leq i < j \leq k, \quad (5.86)$$

and

$$D_{k+1} Q s_i = s_i, \quad 0 \leq i \leq k. \quad (5.87)$$

Let us start by showing that once (5.86) and (5.87) have been proved, we conclude that $D_n = Q^{-1}$. We have assumed that x_0, x_1, \dots, x_{n-1} are nonoptimal, and so p_0, p_1, \dots, p_{n-1} are descent directions (remember, from the last proposition, that D_0, D_1, \dots, D_{n-1} are symmetric and positive definite), and $s_k = \alpha_k p_k$, $k = 0, 1, \dots, n-1$ are nonzero vectors. Since p_0, p_1, \dots, p_{n-1} are Q -conjugate according to (5.86), s_0, s_1, \dots, s_{n-1} will be also Q -conjugate and then, linearly independent by Lemma 5.22. Now, from (5.87) for $k = n-1$, we deduce that $D_n Q$ is the identity matrix, i.e., $D_n Q = I_n$ and so, $D_n = Q^{-1}$.

We start by proving that

$$D_{k+1} Q s_k = s_k, \quad 0 \leq k \leq n-1. \quad (5.88)$$

From $Q s_k = q_k$ and (5.78), we obtain

$$D_{k+1} Q s_k = D_{k+1} q_k = D_k q_k + \frac{s_k (s_k^T q_k)}{s_k^T q_k} - \frac{D_k q_k (q_k^T D_k q_k)}{q_k^T D_k q_k} = s_k.$$

Now we proceed with the proof of (5.86) and (5.87) by induction over k . For $k = 0$, there is nothing to prove in (5.86), while (5.87) holds by (5.88).

Next, we assume that (5.86) and (5.87) are true for k and we shall prove that they are also valid for $k+1$. For $i < k$, one has

$$\nabla f(x_{k+1}) = \nabla f(x_{i+1}) + Q(s_{i+1} + \dots + s_k). \quad (5.89)$$

It happens that s_i is orthogonal to each vector in the right-hand side of (5.89). Obviously, s_i is orthogonal to $Q s_{i+1}, \dots, Q s_k$ since s_0, \dots, s_k are Q -conjugate, and it is also orthogonal to $\nabla f(x_{i+1})$ because the line searches are exact. Thus, we conclude that

$$\nabla f(x_{k+1})^T s_i = 0, \quad 0 \leq i \leq k. \quad (5.90)$$

From (5.90) and (5.87), which is assumed to be true for k , we deduce

$$\nabla f(x_{k+1})^T D_{k+1} Q s_i = \nabla f(x_{k+1})^T s_i = 0, \quad 0 \leq i \leq k. \quad (5.91)$$

Now, as $s_i = \alpha_i p_i$ and $p_{k+1} = -D_{k+1} \nabla f(x_{k+1})$, we have that (5.91) gives rise to

$$0 = \nabla f(x_{k+1})^T D_{k+1} Q s_i = -p_{k+1}^T Q s_i = -\alpha_i p_{k+1}^T Q p_i, \quad 0 \leq i \leq k,$$

i.e.,

$$p_{k+1}^T Q p_i = 0, \quad 0 \leq i \leq k, \quad (5.92)$$

and (5.86) holds for $k + 1$.

Now, from (5.78), we can write

$$D_{k+2} q_i = D_{k+1} q_i + \frac{s_{k+1} s_{k+1}^T}{s_{k+1}^T q_{k+1}} q_i - \frac{D_{k+1} q_{k+1} q_{k+1}^T D_{k+1}}{q_{k+1}^T D_{k+1} q_{k+1}} q_i. \quad (5.93)$$

Since $s_{k+1}^T q_i = s_{k+1}^T Q s_i = 0$, the second term in the right-hand side is zero. Similarly, from (5.87) and (5.92), we deduce

$$q_{k+1}^T D_{k+1} q_i = q_{k+1}^T D_{k+1} Q s_i = q_{k+1}^T s_i = s_{k+1}^T Q s_i = 0,$$

and the third term in the right-hand side of (5.93) also vanishes. Hence, we have that (5.87) and (5.93) lead us to

$$D_{k+2} Q s_i = D_{k+2} q_i = D_{k+1} q_i = D_{k+1} Q s_i = s_i,$$

for $0 \leq i \leq k$. Finally, by (5.88),

$$D_{k+2} Q s_{k+1} = s_{k+1},$$

and we have proved that (5.87) holds for $k + 1$. Then, the proof is finished. \square

5.6 Derivative-Free Optimization Methods

The gradient methods that we have studied in the previous sections require at least calculation of the gradient $\nabla f(x_k)$ and possibly of the Hessian $\nabla^2 f(x_k)$ at each iterate x_k . In many problems, these derivatives either are not explicitly available or are given by complex expressions. For instance, this is the case in chemical engineering, where the evaluation of the objective function may require the execution of an experiment. In such cases, we could use an approximation of the derivatives by finite differences and apply the corresponding gradient method using these approximations. In this

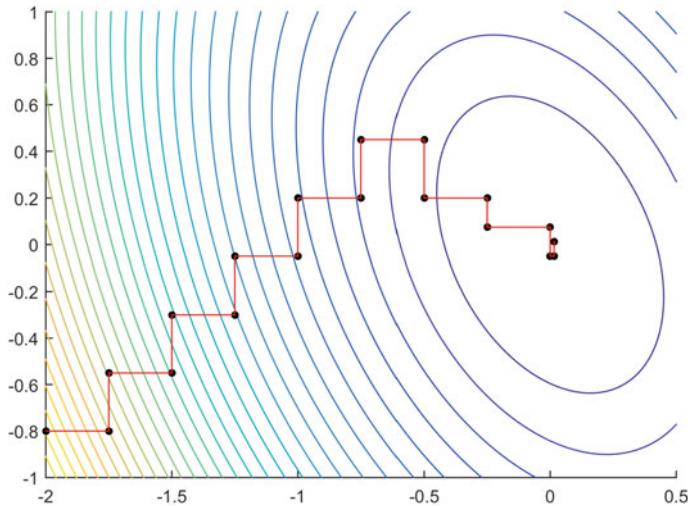


Fig. 5.20 Coordinate descent method

section, we follow a different approach and present other methods that do not use derivatives. More precisely, we focus on *direct search methods*, which are derivative-free methods evaluating the objective function at a finite number of points at each iteration and deciding which actions to take next uniquely based on those function values, without any explicit or implicit derivative considerations.

5.6.1 Coordinate Descent Algorithms

In the coordinate descent method, the objective function is minimized along one coordinate direction in each iteration. The order in which the coordinate directions are chosen may vary during the algorithm. Thus, the method uses one of the coordinate directions e_1, e_2, \dots, e_n (or its opposite directions $-e_i$) as search direction at each step. In the case where the order is cyclical, after n iterations, the method takes again e_1 as search direction. Another variant is the *double sweep method of Aitken* (also called *back-and-forth*), which uses the coordinate directions in the following order

$$e_1, e_2, \dots, e_{n-1}, e_n, e_{n-1}, \dots, e_2, e_1, e_2, \dots$$

These cyclical methods have the advantage of not needing any information about ∇f for choosing the descent directions.

These algorithms are simple and intuitive, but can be quite inefficient. Practical experience shows that typically n iterations of a coordinate descent algorithm are required to get a decrease of the objective function similar to the one achieved by a unique iteration of the steepest descent method.

5.6.2 Nelder and Mead Method

The *Nelder and Mead simplex method* is a direct search algorithm, which is quite different from the line search algorithms we have seen before. To avoid confusion with the simplex method of linear programming, this method is also called the *polytope algorithm*. Each iteration of this method handles a simplex, i.e., the convex hull of $n + 1$ affinely independent points x_0, x_1, \dots, x_n . Let x_{min} and x_{max} be the *best* and the *worst* of the vertices of the simplex, i.e., those vertices satisfying

$$f(x_{min}) = \min_{i=0,1,\dots,n} f(x_i) \quad \text{and} \quad f(x_{max}) = \max_{i=0,1,\dots,n} f(x_i).$$

Let \hat{x} be the centroid (or barycenter) of the simplex facet opposite to x_{max} , i.e.,

$$\hat{x} := \frac{1}{n} \left(-x_{max} + \sum_{i=0}^n x_i \right).$$

The algorithm proceeds as follows.

Algorithm 3: Iteration of the Nelder and Mead simplex method

```

 $x_{ref} = 2\hat{x} - x_{max};$ 
if  $f(x_{min}) > f(x_{ref})$  then
     $x_{exp} = 2x_{ref} - \hat{x};$ 
    if  $f(x_{exp}) < f(x_{ref})$  then
         $| \quad x_{new} = x_{exp}$ 
    else
         $| \quad x_{new} = x_{ref}$ 
    end
else if  $f(x_{min}) \leq f(x_{ref}) < \max_{x_i \neq x_{max}} \{f(x_i)\}$  then
     $| \quad x_{new} = x_{ref}$ 
else
    if  $f(x_{max}) \leq f(x_{ref})$  then
         $| \quad x_{new} = \frac{1}{2}(x_{max} + \hat{x})$ 
    else
         $| \quad x_{new} = \frac{1}{2}(x_{ref} + \hat{x})$ 
    end

```

Case 1: x_{ref} has minimum value
(attempt of expansion)

Case 2: x_{ref} has intermediate value
(use of the reflection)

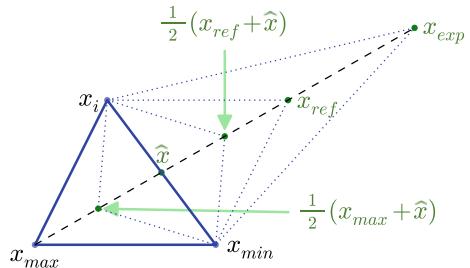
Case 3: x_{ref} has maximum value
(contraction)

Construct the new simplex by replacing x_{max} by x_{new} .

The iteration replaces the *worst* vertex x_{max} by a *better* one. To this aim, the so-called *reflected point*,

$$x_{ref} := 2\hat{x} - x_{max},$$

Fig. 5.21 Possible choices for the new point x_{new} in the simplex method



is computed (this is the point in the straight line determined by x_{max} and \hat{x} which is symmetric to x_{max} with respect to \hat{x}). Depending on the value of the objective function at x_{ref} , with respect to the value of the objective function at the remaining points of the simplex (excluding x_{max}), a new vertex x_{new} is introduced and a new simplex is formed by replacing x_{max} by x_{new} , while keeping the other n vertices (Figs. 5.21 and 5.22).

An important question is how to know when an “appropriate” solution has been found. Nelder and Mead suggested using the standard deviation of the values of the function:

$$test = \sqrt{\frac{1}{n} \sum_{i=0}^n (f(x_i) - M)^2}, \quad \text{where } M = \frac{1}{n+1} \sum_{i=0}^n f(x_i).$$

The algorithm stops when the *test* value is less than a preassigned value of tolerance. This stopping rule turns out to be reasonable in statistical applications where this method is still used. Another possibility would be to stop the algorithm when the value of the function at all the points defining the simplex is the same, that is, when $f(x_{min}) = f(x_{max})$ (or when its difference is less than a certain tolerance value).

When f is not convex, it may be $f(x_{new}) > f(x_{max})$, and there is no improvement of the objective function at the corresponding step. In this case, a possible modification would be to *contract* the simplex toward the best vertex x_{min} , replacing the original vertices x_i by

$$\bar{x}_i = \frac{1}{2}(x_i + x_{min}), \quad i = 0, 1, \dots, n. \quad (5.94)$$

This method, with the described modification, works reasonably well in practice for problems of small size (up to 10), but does not ensure desirable properties of convergence (a counterexample for the convergence with $n = 2$ and f being strictly convex is given by McKinnon [65]).

In Fig. 5.23, we show the result of applying the simplex method to a pair of functions that are often used for testing the performance of algorithms: the Himmelblau function $f(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2$ [51] and the Rosenbrock func-

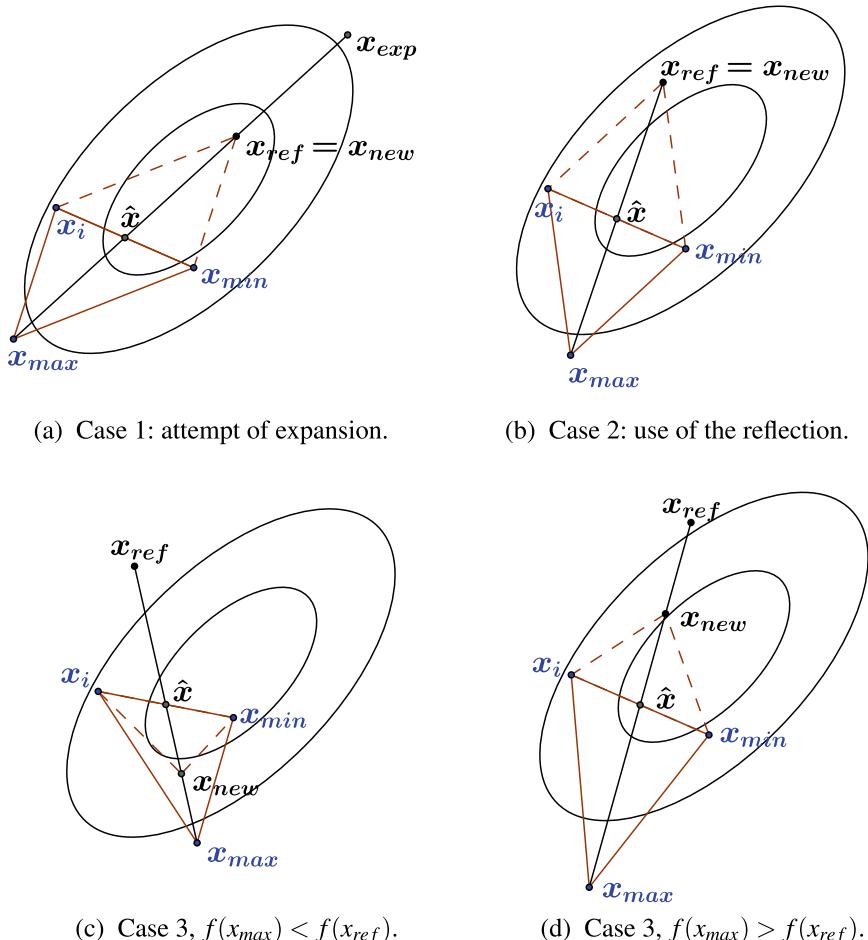


Fig. 5.22 Iterations of the simplex method depending on the level curves of f

tion $f(x, y) = 100(y - x^2)^2 + (1 - x)^2$ [81]. Rosenbrock function is also known as “banana function” because of the shape of its level curves.

5.6.3 Directional Direct Search Methods Using Positive Spanning Sets

Most of the directional direct search methods are based on the use of positive spanning sets and positive bases.

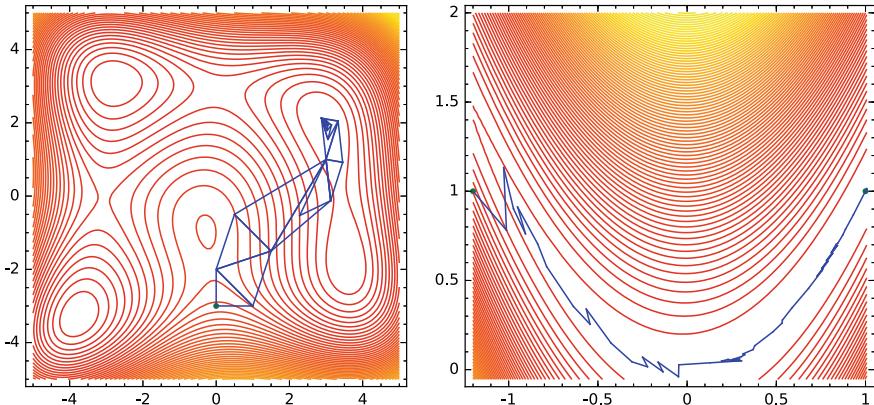


Fig. 5.23 Nelder and Mead simplex method applied to the Himmelblau (left) and the Rosenbrock (right) functions. In blue, we have plotted on the left the simplex generated by the algorithm, while on the right we have only plotted the path determined by x_{min}

Definition 5.27 We call *positive span* of a set of vectors $D = \{d_1, \dots, d_r\} \subset \mathbb{R}^n$ to the convex cone generated by this set, i.e., cone D . A *positive spanning set* in \mathbb{R}^n is a set of vectors whose positive span is \mathbb{R}^n . The set $\{d_1, \dots, d_r\}$ is said to be *positively dependent* if one of the vectors is in the convex cone spanned by the remaining vectors; otherwise, the set is *positively independent*. A *positive basis* in \mathbb{R}^n is a positively independent set whose positive span is \mathbb{R}^n .

Exercise 5.18 asserts that if $\{d_1, \dots, d_r\}$ is a positive spanning set in \mathbb{R}^n , then it contains a subset with $r - 1$ elements that spans \mathbb{R}^n . In some kind of converse statement, let us observe that if $B = \{p_1, \dots, p_n\}$ is an ordinary basis of \mathbb{R}^n , then the set

$$D := \left\{ p_1, \dots, p_n, -\sum_{i=1}^n p_i \right\}, \quad (5.95)$$

is a positive basis of \mathbb{R}^n . Indeed, given any $x \in \mathbb{R}^n$, there are some scalars $\lambda_1, \dots, \lambda_n$ such that

$$x = \lambda_1 p_1 + \lambda_2 p_2 + \dots + \lambda_n p_n.$$

Let $\lambda_{min} := \min\{\lambda_1, \dots, \lambda_n\}$. If $\lambda_{min} \geq 0$, there is nothing to prove. Otherwise, we can write

$$x = (-\lambda_{min}) \left(-\sum_{i=1}^n p_i \right) + (\lambda_1 - \lambda_{min}) p_1 + \dots + (\lambda_n - \lambda_{min}) p_n.$$

The set D in (5.95) is a minimal positive basis in the sense that there is no other positive basis of \mathbb{R}^n with less than $n + 1$ vectors. In particular, in \mathbb{R}^2 the set

$$D_1 = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix} \right\} \quad (5.96)$$

is a minimal positive basis. In [24], different procedures for creating positive basis are presented.

The main motivation to look at a positive spanning set $D \subset \mathbb{R}^n$ is guaranteeing that given any nonzero vector $v \in \mathbb{R}^n$, there is at least one vector $d \in D$ such that v and d form an acute angle, as it is shown next in Lemma 5.28. This has an obvious implication in optimization. Suppose that $f \in C^1(\mathbb{R}^n)$ and $v = -\nabla f(x)$. Then, if D is a positive spanning set, there is a vector $d \in D$ such that $(-\nabla f(x))^T d > 0$ and d is a descent direction. In order to decrease $f(x)$, one only needs to evaluate f at the points $x + \alpha d$ for all $d \in D$, where $\alpha > 0$, and to repeat these evaluations for smaller positive values of α if needed. If the gradient $\nabla f(x)$ is nonzero, there exists a positive value of α and a vector $d \in D$ for which $f(x + \alpha d) < f(x)$. Therefore, this procedure terminates after a finite number of reductions of the parameter α .

Lemma 5.28 *The set $D = \{d_1, \dots, d_r\}$, with $d_i \neq 0_n$, $i = 1, \dots, r$, is a positive spanning set in \mathbb{R}^n if and only if for every $v \in \mathbb{R}^n \setminus \{0_n\}$ there exists an index i in $\{1, \dots, r\}$ such that $v^T d_i > 0$.*

Proof Suppose first that $D = \{d_1, \dots, d_r\}$, with $d_i \neq 0_n$, $i = 1, \dots, r$, is a positive spanning set in \mathbb{R}^n and pick a nonzero vector $v \in \mathbb{R}^n$. Then, there exist nonnegative scalars $\lambda_1, \dots, \lambda_r$ such that $v = \lambda_1 d_1 + \dots + \lambda_r d_r$. Hence,

$$0 < v^T v = \lambda_1 v^T d_1 + \dots + \lambda_r v^T d_r,$$

and necessarily one of the terms, say $\lambda_i v^T d_i$ must be positive, entailing $v^T d_i > 0$.

We prove the converse statement by contradiction. Suppose that for every $v \in \mathbb{R}^n$, there exists an index i in $\{1, \dots, r\}$ such that $v^T d_i > 0$ and cone $D \not\subseteq \mathbb{R}^n$. Since cone D is a closed convex cone with boundary points, by Corollary 2.16, there exists a hyperplane $\{x \in \mathbb{R}^n : c^T x = 0\}$, with $c \neq 0_n$, such that

$$D \subset \text{cone } D \subset \{x \in \mathbb{R}^n : c^T x \leq 0\}. \quad (5.97)$$

Taking $v = c$, (5.97) yields the contradiction. \square

Definition 5.29 The *cosine measure* of a positive spanning set of nonzero vectors D is defined by

$$\text{cm}(D) := \min_{0_n \neq v \in \mathbb{R}^n} \max_{d \in D} \frac{v^T d}{\|v\| \|d\|} = \min_{\|v\|=1} \max_{d \in D} \frac{v^T d}{\|d\|}. \quad (5.98)$$

Lemma 5.28 leads us to conclude that $\text{cm}(D) > 0$ for any positive spanning set of nonzero vectors. For the positive basis in (5.96), $\text{cm}(D_1) = \cos(3\pi/8)$, and if

$$D_2 = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \end{pmatrix} \right\},$$

Fig. 5.24 How far a vector v can be from the vectors in D_1 and D_2

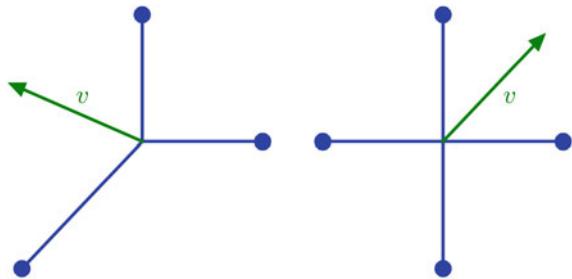
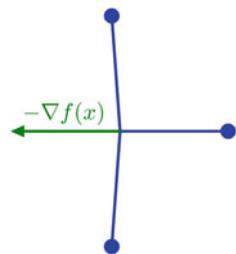


Fig. 5.25 A generating set with a very small cosine measure



$\text{cm}(D_2) = \cos(\pi/4) > \text{cm}(D_1)$ (Fig. 5.24), but D_2 has an element more than D_1 and evaluating f at the points $x + \alpha d$, for all $d \in D_2$, is thus more costly.

Values of $\text{cm}(D)$ close to zero indicate a bad quality of D for numerical purposes, as in the worst case, the angle between $-\nabla f(x)$ and the vectors in D can either be obtuse or nearly perpendicular, as shown in Fig. 5.25.

Obviously, (5.98) implies the following assertion: Given any vector $v \neq 0_n$, there must exist $d \in D$ such that

$$\text{cm}(D)\|v\|\|d\| \leq v^T d. \quad (5.99)$$

Given a positive spanning set D , a point x , and a positive scalar α , we are interested in the points $x + \alpha d$, for all $d \in D$, lying in the ball

$$B := x + \alpha \max_{d \in D} \|d\| \mathbb{B}. \quad (5.100)$$

Theorem 5.30 (Convergence of directional direct search methods) *Let D be a positive spanning set and $\alpha > 0$ be given. Assume that ∇f is Lipschitz continuous, with constant L , in an open set containing the ball B defined in (5.100). If $f(x) \leq f(x + \alpha d)$, for all $d \in D$, then*

$$\|\nabla f(x)\| \leq \frac{\alpha L}{2} \text{cm}(D)^{-1} \max_{d \in D} \|d\|.$$

Proof Consider a vector $d \in D$ such that

$$\text{cm}(D)\|\nabla f(x)\|\|d\| \leq -\nabla f(x)^T d.$$

Now, from Barrow's rule and the fact that $f(x) \leq f(x + \alpha d)$, for all $d \in D$, we obtain

$$0 \leq f(x + \alpha d) - f(x) = \int_0^1 \nabla f(x + t\alpha d)^T (\alpha d) dt.$$

Therefore,

$$\begin{aligned} \text{cm}(D)\|\nabla f(x)\|\|d\|\alpha &\leq \int_0^1 (\nabla f(x + t\alpha d) - \nabla f(x))^T (\alpha d) dt \\ &\leq \int_0^1 \|\nabla f(x + t\alpha d) - \nabla f(x)\| \|\alpha d\| dt \\ &\leq \alpha \|d\| \int_0^1 L \|x + t\alpha d - x\| dt \\ &\leq \frac{L}{2} \alpha^2 \|d\|^2, \end{aligned}$$

and the proof is complete. \square

Theorem 5.30 provides an upper bound on $\|\nabla f(x)\|$ in terms of the size of α times a constant, which depends on the Lipschitz constant of ∇f and the geometry of the positive spanning set D .

A basic directional direct search algorithm iterates as follows.

Algorithm 4: Directional direct-search method

Let D be a positive spanning set.

Choose $x_0, \alpha_0 > 0, \alpha_{tol} > 0, 0 < \beta < 1$ and $\gamma \geq 1$. Set $k = 0$.

while $\alpha_k > \alpha_{tol}$ **do**

then

if $f(x_k + \alpha_k d) < f(x_k)$ for some $d \in D$ **then**
 | $x_{k+1} = x_k + \alpha_k d$
 | $\alpha_{k+1} = \gamma \alpha_k$

else

| $x_{k+1} = x_k$
 | $\alpha_{k+1} = \beta \alpha_k$

end

$k = k + 1$

end

$\left. \begin{array}{l} \text{(successful iteration)} \\ \text{(unsuccessful iteration)} \end{array} \right\}$

$\left. \begin{array}{l} \text{(successful iteration)} \\ \text{(unsuccessful iteration)} \end{array} \right\}$

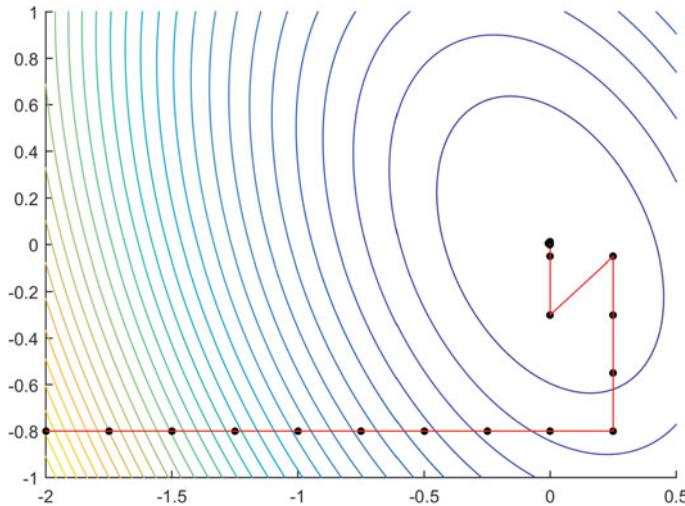


Fig. 5.26 Directional direct search method with positive basis D_1 in (5.96)

In Fig. 5.26, we show the result of applying Algorithm 4 to the same quadratic function used in Fig. 5.20, choosing the same starting point. As explained in Subsection 5.1.2 (see Fig. 5.4), global convergence of such an algorithm cannot be expected in general, unless the condition $f(x_k + \alpha_k d) < f(x_k)$ is replaced by some adequate sufficient decrease condition.

The coordinate descent methods explained in Subsection 5.6.1 are similar to Algorithm 4, except that one changes the positive spanning set at each iteration to $D_k := \{e_i, -e_i\}$, for some $i \in \{1, \dots, n\}$. Note that $\text{cm}(D_k) = 0$, so for $v = -\nabla f(x_k)$, the left-hand side of (5.99) is equal to zero. In fact, these methods can iterate infinitely without reaching a point where the gradient of the function is small, even when exact line searches are used, as shown by Powell [77]. Linear independence of the search directions is not sufficient to guarantee the convergence to a critical point of the functions. This is because the steepest descent direction may become more and more perpendicular to the search direction, and Zoutendijk's condition (5.17) can be satisfied because $\cos \theta_k$ rapidly approaches to zero, even when $\|\nabla f(x_k)\|$ does not. This is not the case for direct search methods using positive spanning sets, for which global convergence results can be proved under some additional assumptions; see [24, Chapter 7].

5.7 Exercises

5.1 Determine the rate of convergence of the following sequences of errors:

- (a) $e(x_k) = \frac{1}{k}$;
- (b) $e(x_k) = (0.5)^{2^k}$;
- (c) $e(x_k) = \frac{1}{k!}$.

5.2 Consider the quadratic function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$

$$f(x) = \frac{1}{2}x^T Qx,$$

where Q is a symmetric matrix of size 2×2 with eigenvalues $0 < \lambda_{\min} < \lambda_{\max}$. Apply the steepest descent method to the problem of minimizing f , with exact line search and initial point

$$x_0 = \frac{1}{\lambda_{\min}}u_{\min} + \frac{1}{\lambda_{\max}}u_{\max},$$

where u_{\min} and u_{\max} are the norm one eigenvectors associated with λ_{\min} and λ_{\max} , respectively.

- (a) Find, by induction, a general expression for the points x_k depending on the parity of k .
- (b) Verify that the bound for $f(x_{k+1})/f(x_k)$ established in (5.28) is attained.
- (c) What property have the even points $x_0, x_2, \dots, x_{2p}, \dots$, and the odd points $x_1, x_3, \dots, x_{2p+1}, \dots$?

5.3 Consider the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ defined by

$$f(x) = x_1^2 + \alpha x_1 x_2 + x_2^2,$$

where α is a positive number.

- (a) Express $f(x)$ in the form $\frac{1}{2}x^T Qx$, with Q being a 2×2 symmetric matrix, and determine the eigenvalues λ_{\min} and λ_{\max} (with $\lambda_{\min} \leq \lambda_{\max}$) of Q , as well as the value of α for which Q is positive definite.
- (b) If u_{\min} and u_{\max} are the associated eigenvectors of norm one, and we take an initial point $x_0 = \frac{1}{\lambda_{\min}}u_{\min} + \frac{1}{\lambda_{\max}}u_{\max}$, what is the value of α for which in three iterations the steepest descent method (with exact line search) obtains a value of the objective function $f(x_3)$ equal to $\frac{f(x_0)}{64}$?

5.4 Consider the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ defined by

$$f(x) = x_1^2 + 3x_1 x_2 + x_2^2.$$

- (a) Express $f(x)$ in the form $\frac{1}{2}x^T Qx$, where Q is a 2×2 symmetric matrix, and determine the eigenvalues λ_{\min} and λ_{\max} (with $\lambda_{\min} \leq \lambda_{\max}$) of Q .
- (b) Let u_{\min} be a nonzero eigenvector associated with λ_{\min} , and take as initial point $x_0 = u_{\min}$. Obtain the expression for the point x_k that results from applying the steepest descent method with constant stepsize ($\alpha_k = \alpha > 0, k = 0, 1, \dots$) to the problem of minimizing $f(x)$ in \mathbb{R}^n . This expression should only depend on α and u_{\min} .
- (c) In view of the expression for x_k obtained in the previous part, to which point does x_k converge when $k \rightarrow \infty$?

5.5 Consider the function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ given by $f(x) = \|x\|^{3/2}$.

- (a) Prove that the gradient is not Lipschitz continuous on any open set U that contains the sublevel set $\{x \in \mathbb{R}^n : f(x) \leq K\}$, where K is any nonnegative constant.
- (b) Suppose that the steepest descent method is applied to minimize the function f , with $\alpha_k = \alpha > 0$, for all k (i.e., with constant stepsize). Prove that, for any α , the method either converges to the unique global optimum $\bar{x} = 0_n$ in a finite number of iterations, or it never converges to this point.

5.6 Consider the function $f : \mathbb{R} \rightarrow \mathbb{R}$ given by

$$f(x) = x^{2p},$$

for an integer p greater or equal than 2.

- (a) Of which type is the convergence of the sequence $\{x_k\}$ generated by the pure Newton's method, to the unique minimizer \bar{x} of the function f over the real line?
- (b) Why quadratic convergence is not achieved? In other words, what hypothesis of Theorem 5.18 (theorem of convergence) fails in this case?

5.7 Suppose that a gradient method is applied in order to minimize the function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, so that $x_{k+1} = x_k + \alpha_k p_k$, $k = 0, 1, 2, \dots$, with

$$p_k := - \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{\partial f(y)}{\partial y_i} \Big|_{y=x_k} \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

where i is the index for which the absolute value of $\frac{\partial f(y)}{\partial y_j} \Big|_{y=x_k}$ is maximized over $j = 1, 2, \dots, n$.

- (a) Prove that p_k is a descent direction and give a negative upper bound for the value of the directional derivative $f'(x_k; p_k)$.
- (b) Assuming that the stepsizes α_k satisfy the Wolfe conditions, and that the function f satisfies the rest of the hypothesis of Zoutendijk's theorem (Theorem 5.6), prove that every limit point of the sequence $\{x_k\}$ is a stationary point.

5.8 Let $f(x) = \frac{1}{2}x^T Qx$, where Q is a symmetric invertible matrix that has at least one negative eigenvalue. Consider the steepest descent method with constant stepsize ($\alpha_k = \alpha > 0$, for all k). Prove that, unless the initial point x_0 belongs to the vector subspace generated by the eigenvectors of Q associated with the positive eigenvalues, the sequence $\{x_k\}$ generated by the algorithm diverges.

5.9 Consider the univariate function

$$f(x) = \frac{2}{3}|x|^3 + \frac{1}{2}x^2,$$

and apply the steepest descent method with stepsize given by $\alpha_k = \frac{\gamma}{k+1}$, where γ is a positive scalar.

(a) Prove that for $\gamma = 1$ and $|x_0| \geq 1$, the method diverges.

(b) Show that the inequality

$$\gamma(2|x_0| + 1) < 2 \quad (5.101)$$

is a sufficient condition for the convergence of the sequence $\{x_k\}$ to the global minimum $\bar{x} = 0$.

5.10 Let $f(x) := \frac{1}{2}x^T Qx - c^T x$ be a quadratic function with Q a symmetric and positive definite matrix. Consider a change of variable $x = Sy$ where S is a diagonal matrix whose i th diagonal element is $q_{ii}^{-1/2}$, and q_{ii} is the i th component of the diagonal of Q .

Prove that if Q is a 2×2 matrix, this “diagonal scaling” improves the conditioning number of the problem and the rate of convergence of the steepest descent method, when applied to the function f .

5.11 Consider the steepest descent method *with error*, and with constant stepsize $\alpha > 0$,

$$x_{k+1} = x_k - \alpha(\nabla f(x_k) + e_k), \quad k = 0, 1, 2, \dots,$$

where e_k is an error such that $\|e_k\| \leq \delta$ for all k , applied to the strictly convex quadratic function

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

Let $q := \max\{|1 - \alpha\lambda_{\min}|, |1 - \alpha\lambda_{\max}|\}$, where λ_{\min} and λ_{\max} are the smallest and largest eigenvalues, respectively, of Q . Prove that if $q < 1$, one has

$$\|x_{k+1}\| \leq \frac{\alpha\delta}{1-q} + q^{k+1}\|x_0\|, \quad \forall k \in \mathbb{N}. \quad (5.102)$$

5.12 Consider the function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined by

$$f(x) = x^3 - 3x + 6.$$

- (a) Find the minima, the maxima, and the points of inflection of f .
- (b) Let \bar{x} be a local minimum of f . Find $\alpha \in]0, 1[$ such that if we take as initial point $x_0 \in]\bar{x} - \alpha, \bar{x} + \alpha[$ and apply the pure Newton's method to the minimization of f , the method will converge to \bar{x} with quadratic rate.
- (c) Taking $x_0 = 0.5$, compute three iterations of the pure Newton's method.

5.13 Consider the function $f :]0, +\infty[\rightarrow \mathbb{R}$ defined by

$$f(x) = x - \ln(x).$$

- (a) Find the minima, maxima, and inflection points (if any) of f .
- (b) Find a positive scalar α such that the second derivative function f'' is Lipschitz continuous on the set $[\alpha, +\infty[$ with Lipschitz constant $L = 16$.
- (c) Let \bar{x} be a minimum of f . Find β such that if we take as initial point $x_0 \in]\bar{x} - \beta, \bar{x} + \beta[$ and apply the pure Newton's method to the minimization of f , the method converges to \bar{x} with quadratic rate.
- (d) Taking $x_0 = 0.1$, $x_0 = 1.5$, and $x_0 = 4$, compute three iterations of the pure Newton's method. Discuss the results obtained.

5.14 One aims to minimize the function

$$f(x) = x - \ln(x - 1) + \ln\left(\frac{19}{3} - x\right)$$

over its domain of existence.

- (a) Find the maxima, minima, and inflection points, if there exist, of f .
- (b) Verify that the second derivative f'' is Lipschitz continuous on the interval $[2, 3]$, and find the Lipschitz constant L in this interval.

Hint: Use Proposition 2.21.

- (c) It is known that the condition

$$\left|1/f''(x)\right| \leq 2\left|1/f''(\bar{x})\right|$$

is verified if $x \in]1, 2.7[$. If \bar{x} is a local minimum of f , find β such that, taking an initial point $x_0 \in]\bar{x} - \beta, \bar{x} + \beta[$ and applying the pure Newton's method to the minimization of f , the method will converge to \bar{x} with quadratic rate.

- (d) Taking $x_0 = 2.5$, compute three iterations of the pure Newton's method. Is x_3 very far from \bar{x} ?

5.15 Consider the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ defined by

$$f(x_1, x_2) = (x_1^2 + x_2^2)^2.$$

- (a) Apply the pure Newton's method to the minimization of this function with initial point $x_0 = (a, a)$, where $a \neq 0$.
- (b) Do we have global convergence in this case?
- (c) If $\{x_k\}$ is the sequence of points generated, which rate of convergence to the global optimum is observed? Why is this rate of convergence slower than the one expected for Newton's method?

5.16 Suppose that the function $f(x) = \frac{1}{2}x^T Qx - b^T x$, where Q is an $n \times n$ symmetric matrix, is strictly convex, and that d_1, d_2, \dots, d_n are Q -conjugate directions. We know that any $x \in \mathbb{R}^n$ admits a unique expression as a linear combination of the form $x = \sum_{i=1}^n \mu_i d_i$. Consider the function $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}$ defined by

$$\varphi(\mu_1, \dots, \mu_n) := f\left(\sum_{i=1}^n \mu_i d_i\right).$$

- (a) Obtain an expression for $\varphi(\mu_1, \dots, \mu_n)$ by making use of the fact that the directions d_1, d_2, \dots, d_n are Q -conjugate.
- (b) Based on the result in (a), argue why it is possible to apply parallel computing, using n processors, to the problem of minimizing f over \mathbb{R}^n .
- (c) Find the optimal value of such a problem, and express it in terms of the values of $d_i^T Q d_i$ and $b^T d_i$, $i = 1, 2, \dots, n$.

5.17 Consider the problem of minimizing over \mathbb{R}^2 the function

$$f(x_1, x_2) = -12x_2 + 4x_1^2 + 4x_2^2 + 4x_1 x_2.$$

- (a) Obtain two Q -conjugate directions, where Q is the Hessian matrix of f .
- (b) Using these Q -conjugate directions, solve the optimization problem taking $x_0 = (-1, 0)^T$ as initial point.
- (c) Apply the conjugate gradient method using $x_0 = (-1, -2)^T$ as initial point.

5.18 Prove that if $\{d_1, \dots, d_r\}$ is a positive spanning set in \mathbb{R}^n , then it contains a subset with $r - 1$ elements that spans \mathbb{R}^n .

5.8 Computer Exercises

In this section, we include five assignments that are designed to help to understand the theory and the algorithms described in this chapter. Each of these assignments can be scheduled for a laboratory session of two hours, and it should be possible to solve them with any mathematical software.

5.8.1 Assignment 1

In this assignment, we will consider the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ given by

$$f(x) = 9x_1^2 - 2x_1 x_2 + x_2^2 - 4x_1^3 + \frac{1}{2}x_1^4.$$

The origin is a global optimum of f that verifies the sufficient second-order optimality conditions. By Taylor's theorem, f must be similar to its second-order approximation nearby the point 0_2 , that is,

$$f(x) \approx f(0_2) + \nabla f(0_2)^T x + \frac{1}{2} x^T \nabla^2 f(0_2) x = \frac{1}{2} x^T \nabla^2 f(0_2) x =: g(x),$$

for every x close to 0_2 . Since $\nabla^2 f(0_2)$ is positive definite, we know by (5.26) that the steepest descent method applied to the quadratic function g verifies

$$\frac{\|x_{k+1}\|}{\|x_k\|} \leq \max\{|1 - \alpha_k \lambda_{\min}|, |1 - \alpha_k \lambda_{\max}|\},$$

where m and M are the smallest and largest eigenvalues of $\nabla^2 f(0_2)$, respectively. We also proved that the value that minimizes this bound is

$$\bar{\alpha}_k = \frac{2}{\lambda_{\min} + \lambda_{\max}}. \quad (5.103)$$

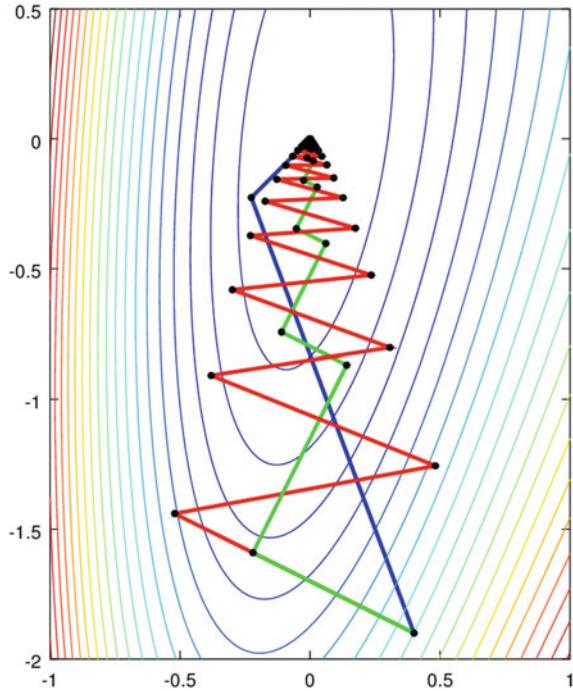
The assignment consists of the following tasks.

1. Compute the gradient, the Hessian matrix, and the critical points of f . Define the function f , a function g with its gradient, and a function H that returns its Hessian matrix, for any vector x . Verify that the origin satisfies the second-order sufficient optimality condition.
2. Make a 3D plot of the function f on $[-1, 5] \times [-3, 8]$.
3. Plot in a different figure 50 level curves of f over the same set.
4. Write a code that runs 40 iterations of the steepest descent method with constant stepsize equal to $\bar{\alpha}_k$ given by (5.103) using as starting point $x_0 = (0.4, -1.9)^T$. Show only $f(x_{40})$ in the screen (For illustration, the iterates are drawn in Fig. 5.27).
5. Write a code that runs 5 iteration of the pure Newton's method with the same starting point used before, showing only the value of $f(x_5)$ in the screen. Compare the result with the value obtained before with the steepest descent method.
6. Compute 40 iterations of the steepest descent method with exact line search (use the appropriate minimization function from your software as a “black box” for finding the stepsize). Compare the result obtained with the ones from the previous two questions.
7. Plot in a new figure the functions $\phi_d(\alpha) := f(x_0 + \alpha p_d)$ and $\phi_N(\alpha) := f(x_0 + \alpha p_N)$ over the interval $[0, 2]$, where $x_0 = (0.4, -1.9)^T$,

$$p_d := -\frac{\nabla f(x_0)}{\|\nabla f(x_0)\|} \quad \text{and} \quad p_N := -\frac{(\nabla^2 f(x_0))^{-1} \nabla f(x_0)}{\|(\nabla^2 f(x_0))^{-1} \nabla f(x_0)\|},$$

correspond to the unitary directions of the steepest descent and Newton's methods, respectively. To differentiate both functions, plot ϕ_d in blue and ϕ_N in red. Add a

Fig. 5.27 We show 40 iterations of the steepest descent method with constant stepsize $\bar{\alpha}_k = \frac{2}{\lambda_{min} + \lambda_{max}}$ (red) and with exact line search (green). We also show 5 iterations of the pure Newton's method (blue)



legend to this plot, with the words “descent” and “Newton.” Answer the following questions (base your response only on the figure obtained):

- Which direction p_d or p_N would be better if we use an exact line search?
- Which direction would be better for very small stepsizes?

5.8.2 Assignment 2

The Rosenbrock function is a nonconvex function that is commonly used to test the performance of optimization algorithms. The function is defined for $(x, y) \in \mathbb{R}^2$ by

$$f(x, y) = 100(y - x^2)^2 + (1 - x)^2.$$

Obviously, $x^* = (1, 1)^T$ is the unique global minimum of the function (in fact, it is the unique critical point of the function).

The assignment consists of the following tasks.

- Make a 3D plot of the Rosenbrock function on $[-2.5, 2.5] \times [-2.5, 4]$ and save the resulting figure as `Pr2_1.png`.
- Plot the level curves of the function on $[-2.5, 2.5] \times [-2.5, 4]$ with level sets at $0, 20, 40, \dots, 400$. Save the plot as `Pr2_2.png`.

3. Create a function named `backtracking` with input parameters `f`, `g`, `xk`, `pk`, `alpha_bar`, `beta`, and `c`. By default, it should take `alpha_bar=1`, `beta=0.5`, `c=1e-4`. The parameter `g` corresponds to the gradient of the function `f`. This function should return a value `alpha` that satisfies the Armijo rule; see Algorithm 2 (Sect. 5.1.2.2).
4. Create a loop that computes 500 iterations of the steepest descent method with backtracking, using as starting point $x_0 = (0, 0.4)^T$. Show only $f(x_0)$, x_{500} and $f(x_{500})$ on the screen. Add to the graph of level curves created in task 2 the result of representing each iteration with a black point and drawing a red line connecting every two consecutive iterations. In the picture, set the limits for the x -axis at $[-0.1, 1]$ and for the y -axis at $[-0.4, 1]$. Save the result as `Pr2_3.png`.
5. Write a loop that computes the number of steps needed to guarantee that $\|x_k - (1, 1)^T\| \leq 10^{-4}$. Create another loop that computes the number of steps needed in order to have $f(x_k) \leq 10^{-4}$.
6. Show the Hessian matrix at the point $(1, 1)^T$, and compute its condition number. Verify if the Hessian matrix is positive definite. What could be the reason for the slow convergence of the steepest descent method?

5.8.3 Assignment 3

In this assignment, we will consider the function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, which is defined for $x \in \mathbb{R}^n$ by

$$f(x) = \sum_{k=1}^n x_k^4,$$

whose global minimum (and unique critical point) is the point 0_n .

Usually, line search methods behave in a similar manner when either the Wolfe conditions or the backtracking method is used to find an adequate stepsize. Nevertheless, we will check in this assignment that the curvature condition required by the Wolfe conditions can sometimes accelerate the convergence.

The assignment consists of the following tasks.

1. Define the function `f` and its gradient `g` in such a way that the definition must be valid for any size of n . Define the function `norm_H` that returns the (Euclidean) norm of its Hessian matrix. Explain why ∇f is Lipschitz continuous on bounded sets. (Hint: Is $\|\nabla^2 f(x)\|$ continuous?)
2. Taking as starting point $x_0 = (-1, 0.4, 0.4, 0.2, 1)^T$, create a loop that runs 15 iterations of the steepest descent method with `backtracking` (which was programmed in Assignment 2). Show on the screen x_{15} and $f(x_{15})$.
3. Create a function named `Wolfe` whose input arguments are `f`, `g`, `x`, `p`, `c1`, and `c2`. By default, `c1=1e-4` and `c2=0.9`. The argument `g` corresponds with the gradient of any function `f`. The function `Wolfe` must be based on Algorithm 1

- (Sect. 5.1.2.1) and must return a stepsize α_k that satisfies the Wolfe conditions.
4. Repeat task 2 choosing now a stepsize that satisfies the Wolfe conditions, using the function created in task 3.
 5. Repeat task 2 changing the search direction: Use the one given by Exercise 5.7, that is,

$$p_k := - \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{\partial f(y)}{\partial y_i} \Big|_{y=x_k} \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

where i is the index for which the absolute value of $\left. \frac{\partial f(y)}{\partial y_j} \right|_{y=x_k}$ is maximized over $j = 1, 2, \dots, n$. First, choose a stepsize using *backtracking*, and then, repeat the experiment requiring the Wolfe conditions.

6. In which of the four methods the best result is obtained?

Remark: Observe that the assumptions of Zoutendijk's theorem (Theorem 5.6) are satisfied in tasks 4 and 5, since:

- p_k is a descent direction, and α_k satisfies the Wolfe conditions.
- $f \in C^1$ is bounded below on \mathbb{R}^5 .
- $S_{f(x_0)} = \{x \in \mathbb{R}^5 : f(x) \leq f(x_0)\} \subset \{x \in \mathbb{R}^5 : |x_k| \leq \sqrt[4]{f(x_0)}\}$, and then $S_{f(x_0)}$ is a bounded set. By task 1, $\|\nabla f\|$ is Lipschitz continuous on $S_{f(x_0)}$.

5.8.4 Assignment 4

In this assignment, we will use the function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ defined by

$$f(x) = -\ln\left(100 - \sum_{i=1}^n x_i\right) - \sum_{i=1}^n \ln(x_i),$$

whose domain is $\{x \in \mathbb{R}_{++}^n : \sum_{i=1}^n x_i < 100\}$.

The assignment consists of the following tasks.

1. Compute the gradient of f , its Hessian matrix and the unique critical point of f . Define the function f , a function g with its gradient, and a function H that returns its Hessian matrix, for any vector x .

2. From now on, take $n = 19$. Define a point x_0 with odd coordinates equal to 1.5 and even coordinates equal to 5.5. This should be done in such a way that if one modifies the value of n , the point x_0 is accordingly updated.
3. Define the critical point of f as xmin in such a way that its value is updated if the value of n is changed. Verify that xmin satisfies the sufficient optimality conditions and, therefore, xmin is a local minimum of f (in fact, it is the global minimum). Is f a convex function?
4. Create a loop that runs 8 iterations of the pure Newton's method using as starting point x_0 . Show on the screen

$$\|x_{k+1} - \text{xmin}\|, \quad \frac{\|x_{k+1} - \text{xmin}\|}{\|x_k - \text{xmin}\|}, \quad \frac{\|x_{k+1} - \text{xmin}\|}{\|x_k - \text{xmin}\|^2}, \quad \text{and} \quad \frac{\|x_{k+1} - \text{xmin}\|}{\|x_k - \text{xmin}\|^3}.$$

Based on the results shown on the screen, does x_k converge to xmin ? If so, is it linearly/quadratically/cubically convergent? Can we apply Theorem 5.18 here?

5. Repeat the experiment in the previous task using now $z_0 = (4.72, 4.72, \dots, 4.72)^T$ as starting point. Answer the same questions of the previous task.
6. Repeat the experiment of task 4 using this time Newton's method with backtracking line search with $\text{alpha_bar}=0.8$ (see Assignment 2), running 20 iterations and using as starting point x_0 defined in task 2. Create a plot showing the number of iterations k on the horizontal axis and the value $\frac{\|x_{k+1} - \text{xmin}\|}{\|x_k - \text{xmin}\|}$ on the vertical axis. Show only the values on $[0, 1]$ for the vertical axis. Does x_k converges linearly/quadratically to xmin ?

5.8.5 Assignment 5

The purpose of this assignment is to code the derivative-free algorithm of Nelder and Mead. To test the algorithm, we will use the Himmelblau function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ defined by

$$f(x, y) = (x^2 + y - 11)^2 + (x + y^2 - 7)^2.$$

This function has a local maximum at $(-0.270845, -0.923039)^T$ with value 181.617 and four global minima at

$$(3, 2)^T, \quad (-2.8051, 3.1313)^T, \quad (-3.7793, -3.2831)^T, \quad (3.5844, -1.8481)^T,$$

where the function is equal to 0.

The assignment consists of the following tasks.

1. Define the function f and the point $x_0 = (0, 1)^T$.
2. Create a function named NM that runs the Nelder and Mead method (see Algorithm 3, Sect. 5.6.2) as follows:

- The input parameters of the function should be: any function f , any starting point x_0 , the number of iterations iter and a constant r whose value is equal to 1 by default.
- The points of the simplex should be stored in a matrix X of size 2×3 , where each column will represent a point of the simplex.
- Create the initial simplex based on x_0 : It will be an equilateral triangle with vertices $x_0 + r(1, 0)^T$, $x_0 + r(-0.5, \sqrt{3}/2)^T$, and $x_0 + r(-0.5, -\sqrt{3}/2)^T$. Store these points in the matrix X .
- Compute the value of the function f at each point of the simplex, and store the result in a row vector named fX , using the same order than in X . Use the appropriate sorting function of your software to sort the vector fX in ascending order, saving the values in $fxord$ and the order in ord .
- Therefore, we will have

$$\begin{aligned}x_{\min} &= X(:, \text{ord}(1)), \quad f(x_{\min}) = fxord(1), \\x_{\max} &= X(:, \text{ord}(3)), \quad f(x_{\max}) = fxord(3),\end{aligned}$$

where $X(:, i)$ stands for the i th column of the matrix X . Define xb as the barycenter of the points different to x_{\max} ; that is, take xb as the middle point between $X(:, \text{ord}(1))$ and $X(:, \text{ord}(2))$.

- Now, we have all the ingredients to write the code for Algorithm 3. Create a loop that runs any number of steps given by iter , without programming the contraction step (5.94) (when $f(x_{\text{new}}) > f(x_{\max})$).
 - Show on the screen the value of X at the end of the loop. The function should return x_{\min} and $f(x_{\min})$.
3. Run the function NM with $\text{iter}=40$, $x_0 = (0, 1)^T$, and verify that the algorithm converges to the point $(3, 2)^T$. Repeat the process using now the point $x_0 = (0, -5)^T$ with 40 iterations and 400 iterations, verifying that the method becomes stagnant and returns a point which is not a minimum. Why do you think this happens? How is the final simplex?
 4. OPTIONAL: Create another function named NM2 that adds to the function NM the contraction step (5.94). This algorithm should also stop when the standard deviation of the values of the function is smaller than a given tolerance tol (which will be an additional input parameter of the function NM2). Verify that NM2 finds the optimal value of the problem for $x_0 = (0, -5)^T$.

Chapter 6

Constrained Optimization



This chapter is devoted to the numerical methods for solving the problem

$$\begin{aligned} P : \text{Min } & f(x) \\ \text{s.t. } & h_j(x) = 0, \quad j = 1, \dots, m, \\ & g_i(x) \leq 0, \quad i = 1, \dots, p, \end{aligned}$$

and their theoretical foundations, where the constraint set C is the whole space \mathbb{R}^n . First, in Section 6.1, the so-called penalty and barrier methods are presented. These methods are based on the idea of approximating constrained optimization problems by unconstrained ones, which can be solved by any of the methods studied in Chapter 5. Both types of methods are driven by a parameter that determines the weight assigned in each iteration to constraint satisfaction relative to minimization of the objective function. In Subsection 6.1.4, a logarithmic barrier approach to linear programming is described as an illustration of the methodology of barrier methods. The subsequent Sections 6.2 to 6.4, of more theoretical flavor, are focused on the formulation of necessary and sufficient optimality conditions, of first and second order, for each of the three types of possible problems: those with equality constraints, with inequality constraints, and with both types of constraints. Conditions of Lagrange, Karush–Kuhn–Tucker, and Fritz John are, respectively, derived through a deep study of the so-called constraint qualifications.

6.1 Penalty and Barrier Methods

Penalty and barrier methods are procedures for approximating constrained optimization problems by unconstrained ones, which could be solved by any of the methods presented in Chapter 5. In the case of penalty methods, this approximation is achieved by the addition of a term to the objective function that penalizes the violation of the

constraint. Barrier methods are obtained by applying a somehow similar technique, with the difference that the corresponding term also adds a high cost to points nearby the boundary of the feasible set, forcing the iterates to lie in its interior. Both methods are driven by a parameter that determines the weight assigned to constraint satisfaction relative to minimization of the objective function. This parameter needs to be increased toward infinity to obtain a good approximation of the original constrained problem. How to do so shall not be discussed here because this is a delicate question, as a large parameter slows down the convergence of most of the algorithms, due to the structure of the resulting unconstrained problem.

6.1.1 Penalty Methods

To motivate the definition of a penalty function, we shall begin by considering two particular types of problems: one with a unique equality constraint and another one with a unique inequality constraint.

Consider first an optimization problem with a unique equality constraint

$$\begin{aligned} P : \text{Min}_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & h(x) = 0. \end{aligned}$$

Suppose that this problem is replaced by the following unconstrained problem, where c is a sufficiently large positive number:

$$P_c : \text{Min}_{x \in \mathbb{R}^n} \{f(x) + c(h(x))^2\}.$$

Intuitively, we see that a reasonable solution \bar{x} to P_c has to be such that $h(\bar{x})$ is close to zero. If not, a small decrease in the value of $h(\bar{x})$ would result in a decrease of the penalty that would offset any possible increase in $f(x)$.

Example 6.1 Consider the following problem

$$\begin{aligned} \text{Min}_{x \in \mathbb{R}} \quad & x \\ \text{s.t.} \quad & -x + 2 \leq 0. \end{aligned}$$

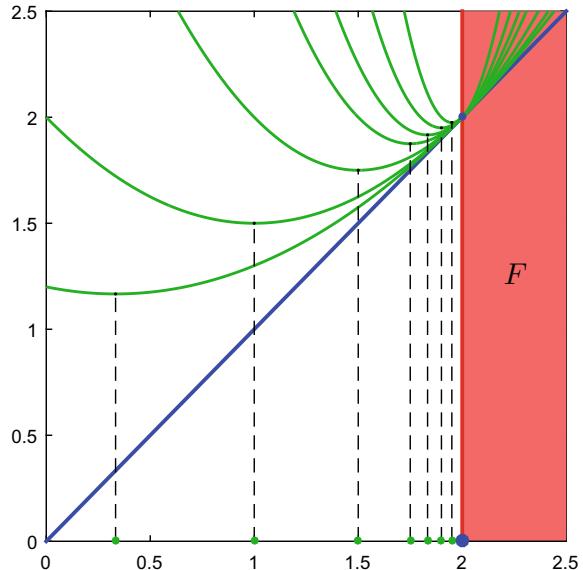
Take $\alpha(x) = (g^+(x))^2$, where $g^+(x) := (p_+ \circ g)(x)$, i.e.,

$$\alpha(x) = \begin{cases} 0, & \text{if } x \geq 2, \\ (-x + 2)^2, & \text{if } x < 2. \end{cases}$$

The minimum of $f + c\alpha$ is reached at $2 - \frac{1}{2c}$, which tends to a minimum of the original problem $\bar{x} = 2$ when $c \rightarrow \infty$; see Fig. 6.1.

Consider now the problem with a single constraint in the form of inequality

Fig. 6.1 The solutions to the penalized problems (in green) converge to the optimal solution of the original problem (in blue) as $c \rightarrow \infty$



$$\begin{aligned} P : \text{Min}_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & g(x) \leq 0. \end{aligned}$$

Clearly, the term $c(g(x))^2$ is not an appropriate penalty since it *punishes* those feasible points that satisfy $g(x) < 0$. A reasonable possibility would be to replace P by the problem

$$\text{Min}_{x \in \mathbb{R}^n} \{f(x) + c \max\{0, g(x)\}\}. \quad (6.1)$$

A difficulty associated with the penalty introduced in (6.1) is that the function $g^+(x) := \max\{0, g(x)\}$ may be nondifferentiable at points x such that $g(x) = 0$. An alternative would be to consider the penalty $c(g^+(x))^2$ whose gradient is $2cg^+(x)\nabla g(x)$, for any $x \in \mathbb{R}^n$.

Overall, an appropriate penalty function has to produce a positive penalty at the infeasible points and no penalty at the feasible points. If the constraints are of the form $h_j(x) = 0$, $j = 1, \dots, m$, $g_i(x) \leq 0$, $i = 1, \dots, p$, then a reasonable *penalty function* would be

$$\alpha(x) := \sum_{j=1}^m \psi(h_j(x)) + \sum_{i=1}^p \phi(g_i(x)), \quad (6.2)$$

where ψ and ϕ are continuous functions satisfying

$$\begin{aligned} \psi(y) &= 0 \text{ if } y = 0, \text{ and } \psi(y) > 0 \text{ if } y \neq 0; \\ \phi(y) &= 0 \text{ if } y \leq 0, \text{ and } \phi(y) > 0 \text{ if } y > 0. \end{aligned} \quad (6.3)$$

For instance, ψ and ϕ can be of the following type:

$$\begin{aligned}\psi(y) &= |y|^{q_1} \\ \phi(y) &= (\max\{0, y\})^{q_2} = (y^+)^{q_2},\end{aligned}$$

where q_1 and q_2 are positive integers. Hence, a typical penalty function is the following

$$\alpha(x) = \sum_{j=1}^m |h_j(x)|^{q_1} + \sum_{i=1}^p (g_i^+(x))^{q_2}.$$

Example 6.2 Let us consider the problem

$$\begin{aligned}\text{Min } &x_1^2 + x_2^2 \\ \text{s.t. } &x_1 + x_2 - 1 = 0.\end{aligned}$$

The unique optimal solution of this problem is $\bar{x} = \left(\frac{1}{2}, \frac{1}{2}\right)^T$, with an associated value of the objective function of $\frac{1}{2}$.

Now, consider the following penalized problem, with $c > 0$,

$$\begin{aligned}\text{Min } &x_1^2 + x_2^2 + c(x_1 + x_2 - 1)^2 \\ \text{s.t. } &x = (x_1, x_2)^T \in \mathbb{R}^2.\end{aligned}$$

Since the objective function of this problem is convex, whatever $c \geq 0$ we take, a necessary and sufficient optimality condition is that its gradient vanishes, i.e.,

$$\begin{aligned}x_1 + c(x_1 + x_2 - 1) &= 0, \\ x_2 + c(x_1 + x_2 - 1) &= 0.\end{aligned}$$

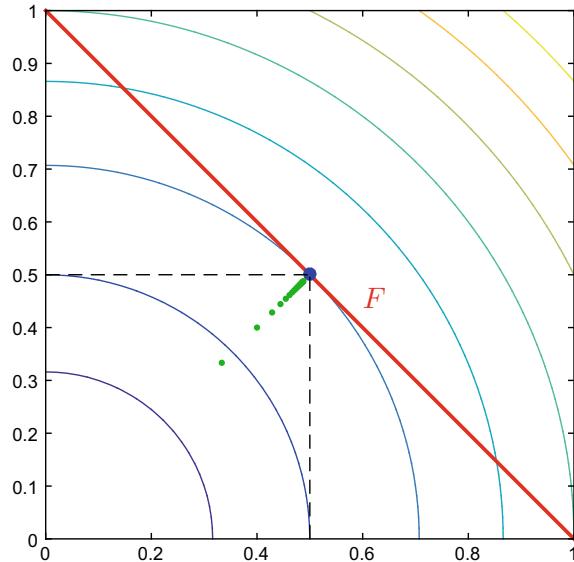
Solving this system, we obtain $x_1 = x_2 = \frac{c}{1+2c}$, and it is evident that the (unique) optimal solution of the penalized problem approaches the optimal solution to the original problem as $c \rightarrow \infty$; see Fig. 6.2.

6.1.2 Methods Using Exterior Penalty Functions

We are dealing again with the problem

$$\begin{aligned}P : \text{Min } &f(x) \\ \text{s.t. } &h(x) = 0_m, \\ &g(x) \leq 0_p,\end{aligned}\tag{6.4}$$

Fig. 6.2 The optimal solutions of the penalized problems (green points) converge to the optimal solution of the original problem (blue point) as $c \rightarrow \infty$



where “ \leq ” denotes the partial ordering in \mathbb{R}^p defined as $u \leq v$ if $u_i \leq v_i$, $i = 1, \dots, p$. So far, we only require that the functions involved (f, h_j, g_i) are continuous. The problem P will be called *primal problem*.

Let α be a continuous function as in (6.2) satisfying the properties (6.3). The basic *exterior penalty method* tries to solve the (*penalty*) *dual problem*

$$\begin{aligned} D_P : \text{Max } & \{\theta(\eta) := \inf_{x \in \mathbb{R}^n} \{f(x) + \eta\alpha(x)\}\} \\ \text{s.t. } & \eta \geq 0. \end{aligned}$$

The optimal value of D_P is $v(D_P) = \sup_{\eta \geq 0} \theta(\eta)$. The fundamental theorem, which is proved below, provides, under certain assumptions, the following double equation result

$$v(D_P) = v(P) = \lim_{\eta \rightarrow \infty} \theta(\eta).$$

The main consequence of this duality result is that the optimum primal value can be approximated, as much as we want, by calculating $\theta(\eta)$ with η sufficiently large. The disadvantage of these procedures is that if x_η is an optimal solution of the inner problem $\theta(\eta)$, i.e.,

$$\theta(\eta) = f(x_\eta) + \eta\alpha(x_\eta), \quad (6.5)$$

it happens that x_η will not generally be feasible for P . For this reason, the functions $\alpha(\cdot)$ are called *exterior* (or *outer*) *penalty functions*.

The fundamental theorem that we have just referred is based on the following lemma.

Lemma 6.3 Let $f, h_1, \dots, h_m, g_1, \dots, g_p$ be continuous functions on \mathbb{R}^n , and let α be a continuous penalty function of the type defined in (6.2) and (6.3). Suppose that for each $\eta > 0$, there exists x_η satisfying (6.5). Then, the following propositions are verified:

- (i) $v(P) \geq v(D_P)$ (weak dual inequality).
- (ii) $f(x_\eta)$ and $\theta(\eta)$ are nondecreasing with η , and $\alpha(x_\eta)$ is nonincreasing with η .

Proof Let $x \in \mathbb{R}^n$ be such that $h(x) = 0_m$ and $g(x) \leq 0_p$. Obviously, $\alpha(x) = 0$, and this yields, whichever $\eta \geq 0$ we take,

$$f(x) = f(x) + \eta\alpha(x) \geq \inf_{y \in \mathbb{R}^n} \{f(y) + \eta\alpha(y)\} = \theta(\eta);$$

hence,

$$f(x) \geq \sup_{\eta \geq 0} \theta(\eta) = v(D_P).$$

Since the last inequality holds for every primal feasible point x , taking the infimum on the primal feasible set, we get $v(P) \geq v(D_P)$; i.e., (i) is proved.

Let us now prove (ii). If $0 < \lambda < \eta$, by the definition of $\theta(\eta)$ and x_η , one has

$$\begin{aligned} f(x_\eta) + \lambda\alpha(x_\eta) &\geq \theta(\lambda) = f(x_\lambda) + \lambda\alpha(x_\lambda), \\ f(x_\lambda) + \eta\alpha(x_\lambda) &\geq \theta(\eta) = f(x_\eta) + \eta\alpha(x_\eta). \end{aligned} \tag{6.6}$$

Summing these inequalities yields

$$(\eta - \lambda)(\alpha(x_\lambda) - \alpha(x_\eta)) \geq 0.$$

Since $\eta > \lambda$, it must be

$$\alpha(x_\lambda) \geq \alpha(x_\eta),$$

and $\alpha(x_\eta)$ certainly is a nonincreasing function of η .

Summing and subtracting $\eta\alpha(x_\eta)$ to the member on the left in (6.6), one gets

$$\theta(\eta) + (\lambda - \eta)\alpha(x_\eta) = f(x_\eta) + \eta\alpha(x_\eta) + (\lambda - \eta)\alpha(x_\eta) \geq \theta(\lambda).$$

Since $\eta > \lambda$ and $\alpha(x_\eta) \geq 0$, it follows that $\theta(\eta) \geq \theta(\lambda)$, and $\theta(\cdot)$ is nondecreasing.

Finally, it remains to prove that $f(x_\eta) \geq f(x_\lambda)$. Otherwise, $f(x_\eta) < f(x_\lambda)$ and

$$f(x_\eta) + \lambda\alpha(x_\eta) < f(x_\lambda) + \lambda\alpha(x_\eta) \leq f(x_\lambda) + \lambda\alpha(x_\lambda),$$

which contradicts (6.6). □

Theorem 6.4 (Convergence of penalty methods) Let P and D_P be the dual pair defined above, and assume that the same conditions as in Lemma 6.3 are verified and that the set $\{x_\eta : \eta \geq 0\}$ is contained in a compact set X . Then:

(i) $v(P) = v(D_P)$ (*dual equality*).

(ii) $v(D_P) = \lim_{\eta \rightarrow \infty} \theta(\eta)$.

(iii) Any accumulation point of the sequence $\{x_{\eta_k}\}$, $k = 1, 2, \dots$, with $\eta_k \rightarrow \infty$, is an optimal solution of P , and $\eta_k \alpha(x_{\eta_k}) \rightarrow 0$ when $k \rightarrow \infty$.

Proof (ii) Since $\theta(\cdot)$ is nondecreasing,

$$v(D_P) = \sup_{\eta \geq 0} \theta(\eta) = \lim_{\eta \rightarrow \infty} \theta(\eta).$$

(i) Let us prove, first, that

$$\lim_{\eta \rightarrow \infty} \alpha(x_\eta) = 0. \quad (6.7)$$

Take y which is a feasible solution of P , and let $\varepsilon > 0$. According to the notation in Lemma 6.3, x_1 is a point such that

$$\theta(1) = f(x_1) + \alpha(x_1).$$

For any η such that

$$\eta \geq \frac{1}{\varepsilon} |f(y) - f(x_1)| + 2,$$

as $\eta \geq 2 > 1$, we have $f(x_\eta) \geq f(x_1)$, by Lemma 6.3(ii). Now, we shall see that $\alpha(x_\eta) < \varepsilon$, which entails $\lim_{\eta \rightarrow \infty} \alpha(x_\eta) = 0$, since $\alpha(x_\eta)$ is nonincreasing with η .

Reasoning by contradiction, if we have $\alpha(x_\eta) \geq \varepsilon$, we could write

$$\begin{aligned} v(P) &\geq v(D_P) \geq \theta(\eta) = f(x_\eta) + \eta \alpha(x_\eta) \\ &\geq f(x_1) + \eta \alpha(x_\eta) \geq f(x_1) + |f(y) - f(x_1)| + 2\varepsilon \\ &\geq f(x_1) + f(y) - f(x_1) + 2\varepsilon > f(y), \end{aligned}$$

but $v(P) > f(y)$ is impossible because y is feasible for P .

Let \bar{x} be an accumulation point of $\{x_{\eta_k}\}$, with $\eta_k \rightarrow \infty$ (it exists by the assumption that this set is contained in a compact set). Without loss of generality, we can assume that $\lim_{k \rightarrow \infty} x_{\eta_k} = \bar{x}$. Thus,

$$v(D_P) = \sup_{\eta \geq 0} \theta(\eta) \geq \theta(\eta_k) = f(x_{\eta_k}) + \eta_k \alpha(x_{\eta_k}) \geq f(x_{\eta_k}).$$

Since $x_{\eta_k} \rightarrow \bar{x}$ and f is continuous, taking limits in the last inequality,

$$v(D_P) \geq \lim_{k \rightarrow \infty} f(x_{\eta_k}) = f(\bar{x}). \quad (6.8)$$

As $\eta_k \rightarrow \infty$, (6.7) leads us to

$$\lim_{k \rightarrow \infty} \alpha(x_{\eta_k}) = 0 = \alpha(\bar{x}).$$

Hence, \bar{x} is a feasible point of P , and (6.8) implies (i).

(iii) Finally, let us observe that

$$\eta_k \alpha(x_{\eta_k}) = \theta(\eta_k) - f(x_{\eta_k}), \quad (6.9)$$

and letting $k \rightarrow \infty$, we have $\lim_{k \rightarrow \infty} \theta(\eta_k) = v(D_P)$, meanwhile $\lim_{k \rightarrow \infty} f(x_{\eta_k}) = f(\bar{x}) = v(P) = v(D_P)$. From (6.9), it follows

$$\lim_{k \rightarrow \infty} \eta_k \alpha(x_{\eta_k}) = 0,$$

and we are done. \square

Corollary 6.5 *If $\alpha(x_\eta) = 0$ for some η , then x_η is optimal for problem P .*

Proof If $\alpha(x_\eta) = 0$, then x_η is feasible for P and, additionally,

$$v(P) \geq \theta(\eta) = f(x_\eta) + \eta \alpha(x_\eta) = f(x_\eta),$$

and this gives rise to the optimality of x_η for P , and $v(P) = v(D_P) = f(x_\eta)$. \square

From Theorem 6.4, it follows that the optimal solution x_η to the problem of minimizing $f(x) + \eta \alpha(x)$, $x \in \mathbb{R}^n$, can be made arbitrarily close to an optimal solution to the original problem simply by taking η large enough. This motivates an algorithmic scheme that consists in solving a sequence of problems of the form

$$\begin{aligned} & \text{Min } f(x) + \eta_k \alpha(x) \\ & \text{s.t. } x \in \mathbb{R}^n, \end{aligned}$$

where $\eta_k \rightarrow +\infty$.

6.1.3 Barrier Methods

The main philosophy in a barrier method is to keep the iterates generated by the algorithm in the interior of the feasible set F . To this aim, a very large cost is imposed on feasible points that lie closer to the boundary $\text{bd}F$, which creates a *barrier* that refrain the iterates from exiting F . For this reason, barrier methods are also referred to as *interior point methods*.

In this section, we confine ourselves to the problem

$$\begin{aligned} P : & \text{Min } f(x) \\ & \text{s.t. } g_i(x) \leq 0, \quad i \in I = \{1, 2, \dots, p\}, \end{aligned} \quad (6.10)$$

where f and g_i , $i \in I$, are continuous functions defined on \mathbb{R}^n , and we assume that P has Slater points; i.e.,

$$S := \{x \in \mathbb{R}^n : g_i(x) < 0, i \in I\} \neq \emptyset.$$

Obviously, $S \subset \text{int } F$. Moreover, if the functions $g_i, i \in I$, are convex and $S \neq \emptyset$, then it can easily be proved that $S = \text{int } F$.

Definition 6.6 A barrier function for P in (6.10) is a function $\beta : \mathbb{R}^n \rightarrow \mathbb{R}$ that satisfies:

- (i) $\beta(x) > 0$ for all $x \in S$.
- (ii) $\beta(x) \rightarrow +\infty$ as $\max_{i \in I} g_i(x) \rightarrow 0$ for $x \in S$.

Two common examples of barrier functions for P are the *logarithmic barrier function*, introduced by Frisch in 1955 [38],

$$\beta(x) := - \sum_{i=1}^p \ln(-g_i(x)),$$

and the *inverse barrier functions*

$$\beta(x) := - \sum_{i=1}^p \frac{1}{g_i(x)^q}, \quad \text{with } q > 0,$$

which were proposed by Carroll for $q = 1$ in 1961 [21].

A *barrier function method* consists in solving a sequence of problems of the form

$$P_B^{\eta_k} : \text{Min}_{x \in S} \left\{ f(x) + \frac{1}{\eta_k} \beta(x) \right\},$$

where $\eta_k \rightarrow +\infty$, with the objective of obtaining a sequence of optimal solutions x_k of $P_B^{\eta_k}$ that approaches an optimal solution of P as $k \rightarrow \infty$. Note that the constraint $x \in S$ in $P_B^{\eta_k}$ is effectively unimportant, as it is never binding.

In Fig. 6.3 we represent the result of applying a logarithmic barrier to the problem

$$\begin{aligned} P : \text{Min } & 4(x_1 - 1)^2 + 2(x_2 - 1)^2 - 4(x_1 - 1)(x_2 - 1) \\ \text{s.t. } & x_2 - \sqrt{1 - x_1^2} \leq 0, \\ & \sqrt{1 - (x_1 - 1)^2} - x_2 + 1 \leq 0. \end{aligned}$$

To this aim, we consider the function $r :]0, +\infty[\times \mathbb{R}^n \rightarrow \mathbb{R}$ defined by

$$r(\eta, x) := f(x) + \frac{1}{\eta} \beta(x). \tag{6.11}$$

The following lemma presents some basic properties of barrier methods.

Lemma 6.7 The function r introduced in (6.11) and any sequence of optimal solutions $x_k, k = 0, 1, 2, \dots$, of the problems $P_B^{\eta_k}, k = 0, 1, 2, \dots$, for positive scalars $\eta_k, k = 0, 1, 2, \dots$, converging to $+\infty$ and strictly increasing, satisfy the following relations, for $k = 0, 1, 2, \dots$:

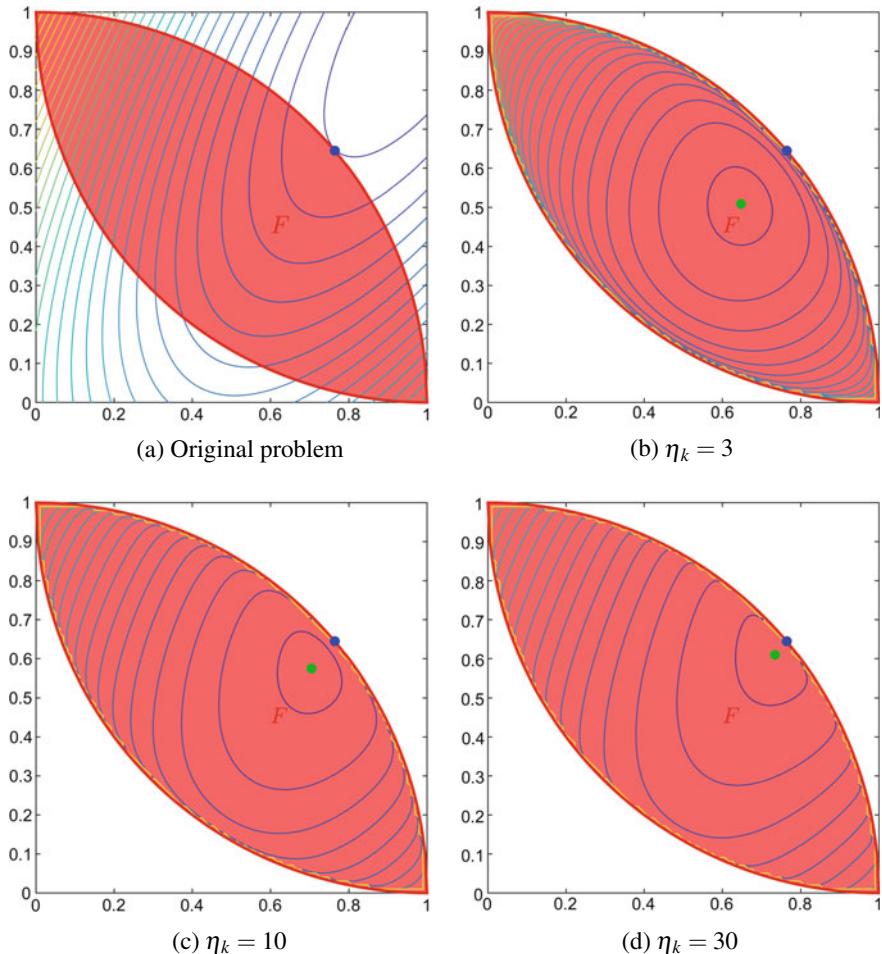


Fig. 6.3 Level curves of the original problem and the logarithmic barrier problems for different values of η_k . The optimal solutions to the barrier problems (in green) approach the optimal solution to the original problem (in blue) as η_k is increased

- (i) $r(\eta_k, x_k) > r(\eta_{k+1}, x_{k+1})$;
- (ii) $\beta(x_k) \leq \beta(x_{k+1})$;
- (iii) $f(x_k) \geq f(x_{k+1})$;
- (iv) if $x^* \in F^*$, then $f(x^*) \leq f(x_k) < r(\eta_k, x_k)$.

Proof (i) As the sequence η_k is strictly increasing, one has

$$\begin{aligned} r(\eta_k, x_k) &= f(x_k) + \frac{1}{\eta_k} \beta(x_k) > f(x_k) + \frac{1}{\eta_{k+1}} \beta(x_k) \\ &\geq f(x_{k+1}) + \frac{1}{\eta_{k+1}} \beta(x_{k+1}) = r(\eta_{k+1}, x_{k+1}). \end{aligned}$$

(ii) Since x_k and x_{k+1} are optimal solutions to $P_B^{\eta_k}$ and $P_B^{\eta_{k+1}}$, respectively, one has

$$f(x_k) + \frac{1}{\eta_k} \beta(x_k) \leq f(x_{k+1}) + \frac{1}{\eta_k} \beta(x_{k+1}),$$

and

$$f(x_{k+1}) + \frac{1}{\eta_{k+1}} \beta(x_{k+1}) \leq f(x_k) + \frac{1}{\eta_{k+1}} \beta(x_k).$$

Summing the latter two inequalities and rearranging, we get

$$\left(\frac{1}{\eta_k} - \frac{1}{\eta_{k+1}} \right) \beta(x_k) \leq \left(\frac{1}{\eta_k} - \frac{1}{\eta_{k+1}} \right) \beta(x_{k+1}),$$

from where it follows that $\beta(x_k) \leq \beta(x_{k+1})$, because $\eta_k < \eta_{k+1}$.

(iii) From (ii) and the optimality of x_{k+1} for $P_B^{\eta_{k+1}}$, we have

$$f(x_k) + \frac{1}{\eta_{k+1}} \beta(x_{k+1}) \geq f(x_k) + \frac{1}{\eta_{k+1}} \beta(x_k) \geq f(x_{k+1}) + \frac{1}{\eta_{k+1}} \beta(x_{k+1}),$$

leading to $f(x_k) \geq f(x_{k+1})$.

(iv) It is obvious, as $f(x^*) \leq f(x_k) < f(x_k) + \frac{1}{\eta_k} \beta(x_k)$. \square

Theorem 6.8 (Barrier convergence theorem) Suppose that f , g_i , $i \in I$, and β are continuous functions. Let $\{x_k\}$ be a sequence of optimal solutions to problems $P_B^{\eta_k}$, $k = 0, 1, 2, \dots$, corresponding to an increasing sequence of positive scalars $\{\eta_k\}$ converging to $+\infty$. Suppose, additionally, that $F^* \cap \text{cl } S \neq \emptyset$. Then, any accumulation point \bar{x} of $\{x_k\}$ is optimal for P , i.e., $\bar{x} \in F^*$.

Proof Let \bar{x} be an accumulation point of $\{x_k\}$, and suppose, for simplicity and without loss of generality, that \bar{x} is in fact the limit of the sequence. From the continuity of f and g_i , $i \in I$, we have that $\lim_{k \rightarrow \infty} f(x_k) = f(\bar{x})$ and $\lim_{k \rightarrow \infty} g_i(x_k) = g_i(\bar{x}) \leq 0$, $i \in I$, as $\{x_k\} \subset S$.

If $x^* \in F^* \cap \text{cl } S \neq \emptyset$, for every $\delta > 0$, we have $(x^* + \delta \mathbb{B}) \cap S \neq \emptyset$. Then, given any $\varepsilon > 0$, and by the continuity of the function f , there exists $\delta_\varepsilon > 0$ such that $f(x) \leq f(x^*) + \varepsilon$ for all $x \in x^* + \delta_\varepsilon \mathbb{B}$. Pick now a point $\tilde{x} \in (x^* + \delta_\varepsilon \mathbb{B}) \cap S$, which thus satisfies $f(\tilde{x}) \leq f(x^*) + \varepsilon$ and $\beta(\tilde{x}) > 0$. Then, for each k ,

$$f(x^*) + \varepsilon + \frac{1}{\eta_k} \beta(\tilde{x}) \geq f(\tilde{x}) + \frac{1}{\eta_k} \beta(\tilde{x}) \geq r(\eta_k, x_k).$$

Since $\beta(\tilde{x}) > 0$ and $\eta_k \rightarrow \infty$, for k sufficiently large,

$$f(x^*) + 2\varepsilon \geq f(x^*) + \varepsilon + \frac{1}{\eta_k} \beta(\tilde{x}) \geq r(\eta_k, x_k).$$

Now, from Lemma 6.7(iv),

$$f(x^*) + 2\varepsilon \geq \lim_{k \rightarrow \infty} r(\eta_k, x_k) \geq f(x^*).$$

As ε can be taken arbitrarily small, we have

$$\lim_{k \rightarrow \infty} r(\eta_k, x_k) = f(x^*).$$

But we also have

$$f(x^*) \leq f(x_k) < r(\eta_k, x_k),$$

so taking limits, we obtain

$$f(x^*) \leq f(\bar{x}) \leq \lim_{k \rightarrow \infty} r(\eta_k, x_k) = f(x^*),$$

whereby $\bar{x} \in F^*$. □

6.1.4 A Logarithmic Barrier Approach to Linear Programming

As an illustration of the methodology of barrier methods, we consider a linear programming problem in standard form,

$$\begin{aligned} P : \text{Min } & c^T x \\ \text{s.t. } & Ax = b, \quad x \geq 0_n, \end{aligned}$$

and for each $\eta > 0$, the associated logarithmic barrier problem,

$$P_B^\eta : \text{Min}_{x \in S} \left\{ c^T x - \frac{1}{\eta} \sum_{i=1}^n \ln x_i \right\}$$

where $x \in \mathbb{R}^n$, $c \in \mathbb{R}^n$, A is an $m \times n$ real matrix, $b \in \mathbb{R}^m$, and

$$S := \{x \in \mathbb{R}^n : Ax = b, \text{ and } x > 0_n\} \neq \emptyset.$$

Problem P_B^η has a strictly convex objective function and linear equality constraints (since the constraints $x > 0_n$ are never binding, they can be neglected). We shall assume that the Slater constraint qualification (SCQ) holds; i.e., that S is nonempty and $\text{rank}(A) = m$. Therefore, problem P_B^η is feasible.

The Lagrangian function of P_B^η is

$$L(x, \mu) = c^T x - \frac{1}{\eta} \sum_{i=1}^n \ln x_i + \mu^T (b - Ax),$$

and the KKT conditions (see Theorem 4.20) for P_B^η are the following:

$$\begin{aligned} c - \frac{1}{\eta} \text{diag}(x)^{-1} 1_n - A^T \mu &= 0_n \\ Ax &= b, \end{aligned} \tag{6.12}$$

where $1_n = (1, 1, \dots, 1)^T \in \mathbb{R}^n$ and $\text{diag}(x)^{-1}$ is a $n \times n$ diagonal matrix whose (i, i) -entry is $1/x_i$.

If, in addition to SCQ, we suppose that the feasible set F is bounded, the problem P_B^η will have a unique optimal solution $x(\eta)$. Defining

$$X := \text{diag}(x),$$

we get an alternative formulation of conditions (6.12)

$$\begin{aligned} Ax &= b, \\ A^T \mu + u &= c, \\ u &= \frac{1}{\eta} X^{-1} 1_n \Leftrightarrow XU 1_n = \frac{1}{\eta} 1_n, \end{aligned} \tag{6.13}$$

where $U = \text{diag}(u)$. Observe also that

$$XU 1_n = \frac{1}{\eta} 1_n \Leftrightarrow x_i u_i = \frac{1}{\eta}, \quad i = 1, 2, \dots, n.$$

Under the current assumptions ($\text{rank}(A) = m$ and S nonempty and bounded), Monteiro and Adler proved in 1989 [66, Propositions 2.1, 2.2, 2.3] that, for each $\eta > 0$, there exists a unique triple $(x(\eta), \mu(\eta), u(\eta))$ verifying (6.13). In particular, $x(\eta)$ is the unique optimal solution of P_B^η . Moreover, they proved that

$$\lim_{\eta \rightarrow \infty} (x(\eta), \mu(\eta), u(\eta)) = (\bar{x}, \bar{\mu}, \bar{u}),$$

where \bar{x} is optimal for P (see Fig. 6.4) and $(\bar{\mu}, \bar{u})$ is optimal for the *linear dual problem*

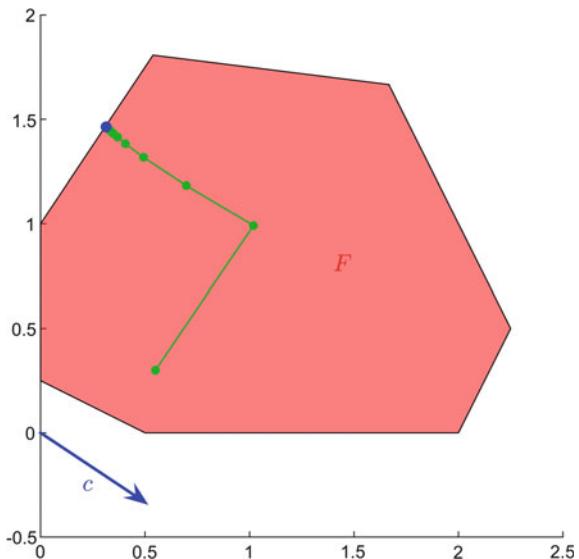
$$\begin{aligned} D : \text{Max } & b^T \mu \\ \text{s.t. } & A^T \mu + u = c, \quad u \geq 0_n. \end{aligned}$$

Let us calculate the duality gap at $(x(\eta), \mu(\eta), u(\eta))$. We have

$$\begin{aligned} c^T x(\eta) - b^T \mu(\eta) &= c^T x(\eta) - (A x(\eta))^T \mu(\eta) \\ &= x(\eta)^T (c - A^T \mu(\eta)) \\ &= x(\eta)^T u(\eta) = \frac{n}{\eta}, \end{aligned}$$

so the duality gap converges to zero when $\eta \rightarrow \infty$.

Fig. 6.4 The solutions to P_B^η (green points) converge to the optimal solution to P (blue point) as $\eta \rightarrow \infty$



For each $\eta > 0$, $(x(\eta), y(\eta), z(\eta))$ is usually obtained by applying the classical Newton method to the KKT system (6.13). This is the so-called (*primal-dual*) *path following method*, which is widely considered to be a notably efficient *interior point method*. The interest in interior point methods comes from the work of Karmarkar [59], where the first linear optimization interior point algorithm with polynomial time complexity was introduced. The reader is addressed to [46] for a comprehensive survey on the field of interior point methods.

Standard convergence results for interior point methods are focused on the behavior of $\|X(\eta)U(\eta)1_n - (1/\eta)1_n\|$, given that $x(\eta)$ and $u(\eta)$ yield the duality gap $x(\eta)^T u(\eta) = n/\eta$. Being a scalar measure associated with two vectors, driving the duality gap to zero is not sufficient to ensure convergence, and some componentwise measure of proximity to the central path must also be sufficiently reduced at each iteration. Monteiro and Adler [66] show that this is ensured, at each iteration, by the condition that $\|X(\eta_k)U(\eta_k)1_n - (1/\eta_k)1_n\| \leq \theta/\eta_k$, for some constant $\theta \in [0, 1/2[$.

6.2 Problems with Equality Constraints

Consider the optimization problem in which the variables are subject only to restrictions in the form of equality

$$\begin{aligned} P : \text{Min } & f(x) \\ \text{s.t. } & h_i(x) = 0, \quad i = 1, \dots, m, \end{aligned} \tag{6.14}$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $h_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, 2, \dots, m$ (equivalently, $h = (h_1, \dots, h_m)^T : \mathbb{R}^n \rightarrow \mathbb{R}^m$). The feasible set is then

$$F := \{x \in \mathbb{R}^n : h(x) = 0_m\} = h^{-1}(0_m).$$

Assume from now on that all the functions (i.e., f and h_i , $i = 1, \dots, m$) are, at least, \mathcal{C}^1 .

We will use the matrix of gradients, i.e., the $n \times m$ matrix whose columns are the gradients of the functions h_i , $i = 1, 2, \dots, m$,

$$\nabla h(x) := [\nabla h_1(x) \mid \dots \mid \nabla h_m(x)],$$

which is the transpose of the Jacobian matrix $J_h(x)$ (recall Definition 1.16).

Theorem 6.9 (Lagrange necessary optimality conditions) *Let \bar{x} be a local minimum of the problem P in (6.14), and assume that the gradients at the point \bar{x} , $\{\nabla h_1(\bar{x}), \dots, \nabla h_m(\bar{x})\}$, are linearly independent (which entails $m \leq n$ and that $\nabla h(\bar{x})$ is full column rank). Then, there exists a unique vector of scalars $\bar{\lambda} = (\bar{\lambda}_1, \dots, \bar{\lambda}_m)^T$ such that:*

$$\nabla f(\bar{x}) + \sum_{i=1}^m \bar{\lambda}_i \nabla h_i(\bar{x}) = \nabla f(\bar{x}) + \nabla h(\bar{x})\bar{\lambda} = 0_n. \quad (6.15)$$

If f and h are \mathcal{C}^2 , we additionally have

$$y^T \left(\nabla^2 f(\bar{x}) + \sum_{i=1}^m \bar{\lambda}_i \nabla^2 h_i(\bar{x}) \right) y \geq 0, \quad \forall y \in V(\bar{x}), \quad (6.16)$$

where

$$V(\bar{x}) = \{y \in \mathbb{R}^n : \nabla h_i(\bar{x})^T y = 0, i = 1, \dots, m\} = \{y \in \mathbb{R}^n : J_h(\bar{x})y = 0_m\}.$$

Remark 6.10 (Comments before the proof) This theorem is called *theorem of Lagrange multipliers*, and the scalars $\bar{\lambda}_1, \bar{\lambda}_2, \dots, \bar{\lambda}_m$ are known as *Lagrange multipliers*. In fact, the system of equations (6.15) is the basis of the so-called *method of Lagrange multipliers*, inspired by the theory developed by this author in 1788, in his book *Mécanique Analytique*. The two most popular proofs of this theorem are based, respectively, one on the implicit function theorem and the second one on the use of a penalization function. We give next the second of these proofs.

Proof • Associated with each k , for $k = 1, 2, \dots$, we consider the function $\Psi_k : \mathbb{R}^n \rightarrow \mathbb{R}$ defined as follows

$$\Psi_k(x) := f(x) + \frac{k}{2} \|h(x)\|^2 + \frac{\alpha}{2} \|x - \bar{x}\|^2,$$

where $\alpha > 0$ is arbitrary.

Since \bar{x} is a local minimum of the problem P , there exists $\varepsilon > 0$ such that $f(\bar{x}) \leq f(x)$ for all $x \in F \cap (\bar{x} + \varepsilon\mathbb{B})$. For each k , we take a point

$$x_k \in \operatorname{argmin}_{x \in \bar{x} + \varepsilon\mathbb{B}} \Psi_k(x).$$

(Such a point exists by the Weierstrass theorem, due to the continuity of Ψ_k and the compactness of $\bar{x} + \varepsilon\mathbb{B}$.) Optimality of x_k implies

$$\Psi_k(x_k) = f(x_k) + \frac{k}{2} \|h(x_k)\|^2 + \frac{\alpha}{2} \|x_k - \bar{x}\|^2 \leq \Psi_k(\bar{x}) = f(\bar{x}). \quad (6.17)$$

- Since $\{x_k\} \subset \bar{x} + \varepsilon\mathbb{B}$, there will be at least an accumulation point u of the sequence; i.e., there will be a subsequence $\{x_{k_r}\}$ converging to u (with $u \in \bar{x} + \varepsilon\mathbb{B}$). Now, we prove that

$$h(u) = \lim_{r \rightarrow \infty} h(x_{k_r}) = 0_m,$$

that is, $u \in F$. Otherwise, if

$$\lim_{r \rightarrow \infty} \|h(x_{k_r})\| = \|h(u)\| > 0,$$

by taking limits in (6.17) as $r \rightarrow \infty$, we would get a contradiction, since

$$\lim_{r \rightarrow \infty} \left\{ f(x_{k_r}) + \frac{\alpha}{2} \|x_{k_r} - \bar{x}\|^2 \right\} = f(u) + \frac{\alpha}{2} \|u - \bar{x}\|^2,$$

whereas

$$\lim_{r \rightarrow \infty} \frac{k_r}{2} \|h(x_{k_r})\|^2 = +\infty,$$

which yields the following contradiction

$$\lim_{r \rightarrow \infty} \left\{ f(x_{k_r}) + \frac{k_r}{2} \|h(x_{k_r})\|^2 + \frac{\alpha}{2} \|x_{k_r} - \bar{x}\|^2 \right\} = +\infty \leq f(\bar{x}).$$

- Inequality (6.17) entails

$$f(x_{k_r}) + \frac{\alpha}{2} \|x_{k_r} - \bar{x}\|^2 \leq f(\bar{x}),$$

and taking again limits as $r \rightarrow \infty$, we get

$$f(u) + \frac{\alpha}{2} \|u - \bar{x}\|^2 \leq f(\bar{x}).$$

Since $f(\bar{x}) \leq f(u)$ as $u \in F \cap (\bar{x} + \varepsilon\mathbb{B})$, we conclude $\|u - \bar{x}\| = 0$, i.e., $u = \bar{x}$. Therefore, \bar{x} turns out to be the unique accumulation point of the sequence $\{x_k\}$, implying that

$$\lim_{k \rightarrow \infty} x_k = \bar{x}.$$

- The convergence of x_k to \bar{x} entails that, for k large enough, x_k is an interior point of $\bar{x} + \varepsilon\mathbb{B}$ and so, x_k is an unrestricted local minimum of $\Psi_k(\cdot)$. Applying then the necessary optimality condition of first order for unrestricted local minima, we deduce

$$0_n = \nabla \Psi_k(x_k) = \nabla f(x_k) + k \nabla h(x_k)h(x_k) + \alpha(x_k - \bar{x}). \quad (6.18)$$

Since $\nabla h(\bar{x})$ is full rank, $\nabla h(x_k)$ is also full rank if k is large enough (remember that $h_i \in \mathcal{C}^1$, $i = 1, 2, \dots, m$), and

$$\nabla h(x_k)^T \nabla h(x_k)$$

is a nonsingular $m \times m$ matrix. This allows us to multiply (6.18) by

$$(\nabla h(x_k)^T \nabla h(x_k))^{-1} \nabla h(x_k)^T,$$

yielding

$$kh(x_k) = -(\nabla h(x_k)^T \nabla h(x_k))^{-1} \nabla h(x_k)^T (\nabla f(x_k) + \alpha(x_k - \bar{x})).$$

Taking once more limits when $k \rightarrow \infty$, we observe that the sequence of vectors $\{kh(x_k)\}$ converges to

$$\bar{\lambda} := -(\nabla h(\bar{x})^T \nabla h(\bar{x}))^{-1} \nabla h(\bar{x})^T \nabla f(\bar{x}).$$

If we let $k \rightarrow \infty$ in (6.18), we finally obtain

$$0_n = \nabla f(\bar{x}) + \nabla h(\bar{x})\bar{\lambda},$$

which is (6.15).

- Simple calculations provide

$$\frac{\partial \Psi_k(x)}{\partial x_j} = \frac{\partial f(x)}{\partial x_j} + k \sum_{p=1}^m h_p(x) \frac{\partial h_p(x)}{\partial x_j} + \alpha(x_j - \bar{x}_j),$$

and

$$\frac{\partial^2 \Psi_k(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j} + k \left(\sum_{p=1}^m \frac{\partial h_p(x)}{\partial x_i} \frac{\partial h_p(x)}{\partial x_j} + \sum_{p=1}^m h_p(x) \frac{\partial^2 h_p(x)}{\partial x_i \partial x_j} \right) + \alpha \delta_{ij},$$

where

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

The second-order necessary optimality condition for unrestricted local minima establishes that, for k large enough, the Hessian matrix

$$\nabla^2 \Psi_k(x_k) = \nabla^2 f(x_k) + k \nabla h(x_k) \nabla h(x_k)^T + k \sum_{i=1}^m h_i(x_k) \nabla^2 h_i(x_k) + \alpha I$$

is positive semidefinite, whichever α we take.

Let us take a fixed $y \in V(\bar{x})$ (i.e., $\nabla h(\bar{x})^T y = 0_m$). Remembering that, for k large enough, the matrix $\nabla h(x_k)^T \nabla h(x_k)$ is invertible, we can easily verify that

$$y_k := y - \nabla h(x_k) (\nabla h(x_k)^T \nabla h(x_k))^{-1} \nabla h(x_k)^T y \in V(x_k). \quad (6.19)$$

Since $\nabla h(x_k)^T y_k = 0_m$ and the matrix $\nabla^2 \Psi_k(x_k)$ is positive semidefinite, one has

$$0 \leq y_k^T \nabla^2 \Psi_k(x_k) y_k = y_k^T \left(\nabla^2 f(x_k) + k \sum_{i=1}^m h_i(x_k) \nabla^2 h_i(x_k) \right) y_k + \alpha \|y_k\|^2. \quad (6.20)$$

Due to the facts that $\nabla h(\bar{x})^T y = 0_m$ and $x_k \rightarrow \bar{x}$, we deduce from (6.19) that $y_k \rightarrow y$. Taking limits in (6.20) and considering that $kh_i(x_k) \rightarrow \bar{\lambda}_i$, as $k \rightarrow \infty$, we get

$$0 \leq y^T \left(\nabla^2 f(\bar{x}) + \sum_{i=1}^m \bar{\lambda}_i \nabla^2 h_i(\bar{x}) \right) y + \alpha \|y\|^2.$$

Since α can be taken arbitrarily close to zero, we actually obtain

$$0 \leq y^T \left(\nabla^2 f(\bar{x}) + \sum_{i=1}^m \bar{\lambda}_i \nabla^2 h_i(\bar{x}) \right) y.$$

As y is an arbitrary element of $V(\bar{x})$, we are done. □

Example 6.11 Consider the optimization problem in \mathbb{R}^2

$$\begin{aligned} P : \text{Min } f(x) &= x_1^2 + x_2^2 - x_1 x_2 + x_1 \\ \text{s.t. } h_1(x) &= (1 - x_1)^3 - x_2 = 0, \end{aligned}$$

which is represented in Fig. 6.5. The objective function is quadratic and coercive (recall Proposition 3.1), so it has a global minimum by Theorem 1.38.

We know by Theorem 6.9 that every candidate for local minimum must satisfy the first-order optimality condition (6.15):

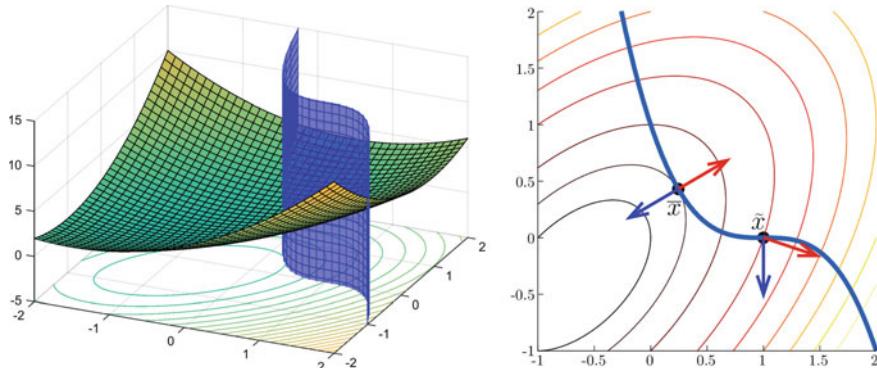


Fig. 6.5 The point $\tilde{x} = (1, 0)^T$ cannot be a candidate for a local minimum. At the global minimum \bar{x} , Theorem 6.9 clearly holds

$$\nabla f(x) + \lambda \nabla h(x) = \begin{pmatrix} 2x_1 - x_2 + 1 - 3\lambda(1 - x_1)^2 \\ 2x_2 - x_1 - \lambda \end{pmatrix} = 0_2. \quad (6.21)$$

Observe, for instance, that the feasible point $\tilde{x} = (1, 0)^T$ cannot be a candidate for a local minimum, since

$$\nabla f(\tilde{x}) + \lambda \nabla h(\tilde{x}) = \begin{pmatrix} 3 \\ -1 - \lambda \end{pmatrix} \neq 0_2, \quad \forall \lambda \in \mathbb{R}.$$

By (6.21) and the feasibility condition $x_2 = (1 - x_1)^3$, we deduce after some algebra that the global minimum must satisfy the equation

$$6x_1^5 - 30x_1^4 + 64x_1^3 - 69x_1^2 + 38x_1 - 6 = 0. \quad (6.22)$$

Thanks to the fundamental theorem of algebra (Theorem 1.44), we know that the latter polynomial of degree five has exactly five roots on \mathbb{C} . As complex roots come in pairs, it must have at least one real root, which can be numerically found, e.g., using Newton's method with starting point 0 (see Section 5.4). This gives us the approximation to the global minimum

$$\bar{x} = (0.2445258, 0.4311803)^T.$$

In fact, using Sturm's theorem [83, Theorem 5.6.2], one can easily show that the polynomial (6.22) has exactly one real root, and therefore, this real root must be the abscissa of the global minimum of problem P .

Observe in Fig. 6.5 that the first-order necessary condition clearly holds at \bar{x} , while this is not the case for \tilde{x} .

The following example illustrates the situation in which the gradients at \bar{x} , $\nabla h_1(\bar{x}), \dots, \nabla h_m(\bar{x})$, are not linearly independent.

Example 6.12 Consider the optimization problem in \mathbb{R}^2

$$\begin{aligned} P : \text{Min } f(x) &= x_1 + x_2 \\ \text{s.t. } h_1(x) &= (x_1 - 1)^2 + x_2^2 - 1 = 0, \\ h_2(x) &= (x_1 - 2)^2 + x_2^2 - 4 = 0. \end{aligned}$$

Observe that at the minimum $\bar{x} = (0, 0)^T$ the gradient of the objective function, $\nabla f(\bar{x}) = (1, 1)^T$, cannot be expressed as a linear combination of the gradients $\nabla h_1(\bar{x}) = (-2, 0)^T$ and $\nabla h_2(\bar{x}) = (-4, 0)^T$. Thus, the first-order necessary optimality condition (6.15) is not satisfied for any $\bar{\lambda}_1$ and $\bar{\lambda}_2$.

Very often, the optimality conditions above are written in terms of the *Lagrange function* of problem P , which now is the function $L : \mathbb{R}^{n+m} \rightarrow \mathbb{R}$ defined by

$$L(x, \lambda) := f(x) + \sum_{i=1}^m \lambda_i h_i(x).$$

Then, the necessary optimality conditions (6.15) and (6.16), together with the feasibility condition $h(\bar{x}) = 0_m$, can be expressed in the following simple way involving the Lagrange function and its derivatives

$$\nabla_x L(\bar{x}, \bar{\lambda}) = 0_n, \quad \nabla_\lambda L(\bar{x}, \bar{\lambda}) = 0_m, \quad (6.23)$$

$$y^T \nabla_{xx}^2 L(\bar{x}, \bar{\lambda}) y \geq 0, \quad \forall y \in V(\bar{x}). \quad (6.24)$$

As it happens in the unconstrained case, a solution of the system ($n + m$ equations with $n + m$ unknowns) (6.23) could even correspond to a maximum. This is the case in the following example.

Example 6.13 Let us consider the problem in \mathbb{R}^3

$$\begin{aligned} P : \text{Min } &\frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \\ \text{s.t. } &x_1 + x_2 + x_3 = 3. \end{aligned}$$

The first-order necessary optimality conditions (6.23) yield the following system

$$\begin{aligned} \bar{x}_1 + \bar{\lambda} &= 0, \\ \bar{x}_2 + \bar{\lambda} &= 0, \\ \bar{x}_3 + \bar{\lambda} &= 0, \\ \bar{x}_1 + \bar{x}_2 + \bar{x}_3 &= 3. \end{aligned}$$

This is a system of four equations and four unknowns ($n + m = 3 + 1 = 4$), having a unique solution

$$\bar{x}_1 = \bar{x}_2 = \bar{x}_3 = 1, \quad \bar{\lambda} = -1.$$

The gradient of h is constant and equal to $(1, 1, 1)^T$, and every feasible point satisfies the condition of linear independence of the gradients of functions h_i . Therefore, $\bar{x} = (1, 1, 1)^T$ is the unique candidate to be a local minimum. Moreover, $\nabla_{xx}^2 L(\bar{x}, \bar{\lambda})$ is the identity matrix and the second-order necessary optimality condition is trivially satisfied. In this way, the point $\bar{x} = (1, 1, 1)^T$ is certainly credited as the sole candidate for local minimum.

To make a final decision on whether \bar{x} is certainly a local minimum, we need to apply some sufficient conditions of optimality, although in this case we can also appeal to a simple perturbation argument by which it is immediately verified that $\bar{x} = (1, 1, 1)^T$ is a local minimum of the function f on $\{x : h(x) = 0\}$ (and therefore, it is also global minimum by the convexity of f).

Let $z = (z_1, z_2, z_3)^T \neq 0_3$ be such that $h(\bar{x} + z) = 0$ (i.e., z is a variation vector preserving feasibility). Therefore,

$$(\bar{x}_1 + z_1) + (\bar{x}_2 + z_2) + (\bar{x}_3 + z_3) = 3 \Rightarrow z_1 + z_2 + z_3 = 0,$$

which entails

$$\begin{aligned} f(\bar{x} + z) &= \frac{1}{2}[(\bar{x}_1 + z_1)^2 + (\bar{x}_2 + z_2)^2 + (\bar{x}_3 + z_3)^2] \\ &= f(\bar{x}) + (z_1 + z_2 + z_3) + \frac{1}{2}(z_1^2 + z_2^2 + z_3^2) \\ &= f(\bar{x}) + \frac{1}{2}(z_1^2 + z_2^2 + z_3^2) > f(\bar{x}). \end{aligned}$$

Now suppose that, instead of considering the problem P above, we were faced with the problem

$$\begin{aligned} \tilde{P} : \text{Min } &- \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \\ \text{s.t. } &x_1 + x_2 + x_3 = 3. \end{aligned}$$

Conditions (6.23) applied to \tilde{P} give rise to

$$\bar{x} = (1, 1, 1)^T \quad \text{and} \quad \bar{\lambda} = 1.$$

However, the second-order necessary optimality condition (6.24) is not fulfilled, and since every feasible point satisfies the linear independence condition, there is no local minimum for problem \tilde{P} .

Next, we establish sufficient conditions of optimality for the optimization problem with equality constraints. The proof of the corresponding theorem is based on the following lemma.

Lemma 6.14 Let A and B be two symmetric matrices $n \times n$. Assume that B is positive semidefinite, while A is positive definite on the null space of B , i.e.,

$$x^T Ax > 0, \quad \forall x \neq 0_n \text{ such that } Bx = 0_n. \quad (6.25)$$

Then, there is a scalar \bar{c} such that $A + cB$ is positive definite for all $c \geq \bar{c}$.

Proof If there exists some $\bar{c} \in \mathbb{R}$ such that $A + \bar{c}B$ is positive definite, then, for all $x \neq 0_n$ and all $c \geq \bar{c}$, we have that

$$0 < x^T (A + \bar{c}B)x \leq x^T Ax + cx^T Bx = x^T (A + cB)x,$$

thanks to the positive semidefiniteness of B . Thus, the claim holds.

Otherwise, assume that there is no $\bar{c} \in \mathbb{R}$ such that $A + \bar{c}B$ is positive definite. Then, for each $k = 1, 2, \dots$, there exists x_k with $\|x_k\| = 1$ such that

$$x_k^T Ax_k + kx_k^T Bx_k \leq 0. \quad (6.26)$$

Since $\{x_k\}$ is contained in a compact set, there exists a subsequence $\{x_{k_r}\}$ which converges to some \bar{x} (with $\|\bar{x}\| = 1$). Taking upper limits in (6.26) with $k = k_r$ and $r \rightarrow \infty$, we get

$$\bar{x}^T A\bar{x} + \limsup_{r \rightarrow \infty} (k_r x_{k_r}^T Bx_{k_r}) \leq 0. \quad (6.27)$$

Since B is positive semidefinite, we have that $x_{k_r}^T Bx_{k_r} \geq 0$, so the sequence $\{x_{k_r}^T Bx_{k_r}\}$ must converge to zero (otherwise, the left-hand side term in (6.27) would be $+\infty$). Further, we deduce from (6.27) that $\bar{x}^T A\bar{x} \leq 0$.

Therefore, we have proved that $\bar{x}^T B\bar{x} = 0$, which implies $B\bar{x} = 0_n$ and, by (6.25) that $\bar{x}^T A\bar{x} > 0$, a contradiction. \square

The proof of the sufficient optimality conditions for the problem P is based on the notion of *augmented Lagrangian*, which is the function

$$L_c(x, \lambda) := f(x) + \lambda^T h(x) + \frac{c}{2} \|h(x)\|^2,$$

with $c \in \mathbb{R}$. The augmented Lagrangian is nothing else but the ordinary Lagrangian function of the problem

$$\begin{aligned} P_c : \text{Min } & f(x) + \frac{c}{2} \|h(x)\|^2 \\ \text{s.t. } & h(x) = 0_m, \end{aligned}$$

which has the same local minima that the original problem of minimizing $f(x)$ subject to $h(x) = 0_m$. The gradient and the Hessian matrix of L_c with respect to x are

$$\begin{aligned} \nabla_x L_c(x, \lambda) &= \nabla f(x) + \nabla h(x)(\lambda + ch(x)), \\ \nabla_{xx}^2 L_c(x, \lambda) &= \nabla^2 f(x) + \sum_{i=1}^m (\lambda_i + ch_i(x)) \nabla^2 h_i(x) + c \nabla h(x) \nabla h(x)^T. \end{aligned}$$

Theorem 6.15 (Sufficient optimality conditions) Assume that f and h_i , $i = 1, \dots, m$, are \mathcal{C}^2 . Suppose that $\bar{x} \in \mathbb{R}^n$ and $\bar{\lambda} \in \mathbb{R}^m$ satisfy (6.23) and the following condition:

$$y^T \nabla_{xx}^2 L(\bar{x}, \bar{\lambda}) y > 0, \quad \forall y \neq 0_n \text{ such that } \nabla h(\bar{x})^T y = 0_m. \quad (6.28)$$

Then, \bar{x} is a strict local minimum of problem P . Moreover, there exist scalars $\gamma > 0$ and $\varepsilon > 0$ such that

$$f(x) \geq f(\bar{x}) + \frac{\gamma}{2} \|x - \bar{x}\|^2, \quad \forall x \text{ such that } h(x) = 0_m \text{ and } \|x - \bar{x}\| < \varepsilon.$$

Proof If \bar{x} and $\bar{\lambda}$ satisfy (6.23), we have

$$\nabla_x L_c(\bar{x}, \bar{\lambda}) = \nabla f(\bar{x}) + \nabla h(\bar{x})(\bar{\lambda} + c h(\bar{x})) = \nabla_x L(\bar{x}, \bar{\lambda}) = 0_n, \quad (6.29)$$

$$\nabla_{xx}^2 L_c(\bar{x}, \bar{\lambda}) = \nabla_{xx}^2 L(\bar{x}, \bar{\lambda}) + c \nabla h(\bar{x}) \nabla h(\bar{x})^T. \quad (6.30)$$

By (6.28), $y^T \nabla_{xx}^2 L(\bar{x}, \bar{\lambda}) y > 0$ for all y such that $\nabla h(\bar{x})^T y = 0_m$ (equivalently, for all y belonging to the null space of $\nabla h(\bar{x}) \nabla h(\bar{x})^T$). Applying Lemma 6.14, there exists \bar{c} such that, by (6.30),

$$\nabla_{xx}^2 L_c(\bar{x}, \bar{\lambda}) \text{ is positive definite } \forall c > \bar{c}. \quad (6.31)$$

Next, we proceed by applying the sufficient optimality conditions to the unrestricted optimization problem (Theorem 1.27). They lead us to conclude, by (6.29) and (6.31), that, for all $c > \bar{c}$, \bar{x} is a strict local minimum of the function $L_c(\cdot, \bar{\lambda})$ and, in addition, there exist $\gamma > 0$ and $\varepsilon > 0$ such that

$$L_c(x, \bar{\lambda}) \geq L_c(\bar{x}, \bar{\lambda}) + \frac{\gamma}{2} \|x - \bar{x}\|^2,$$

for all x such that $\|x - \bar{x}\| < \varepsilon$. Since for all x verifying $h(x) = 0$, we have $L_c(x, \bar{\lambda}) = f(x)$, it follows that

$$f(x) \geq f(\bar{x}) + \frac{\gamma}{2} \|x - \bar{x}\|^2,$$

for all x such that $h(x) = 0_m$ and $\|x - \bar{x}\| < \varepsilon$. The proof is complete. \square

Example 6.16 To illustrate the application of the last theorem, we shall consider the following problem:

$$\begin{aligned} P : \text{Min } f(x) &= \frac{1}{2}(x_1^2 - x_2^2) - x_2, \\ \text{s.t. } x_2 &= 0. \end{aligned}$$

It can be easily checked that $\bar{x} = (0, 0)^T$ and $\bar{\lambda} = 1$ form the unique pair (x, λ) satisfying conditions (6.23) and (6.28). Obviously, $\bar{x} = (0, 0)^T$ is the unique global minimum of problem P (which is equivalent to minimize $\frac{1}{2}x_1^2$ on \mathbb{R} , and take $\bar{x}_2 = 0$).

The augmented Lagrangian is

$$L_c(x, \bar{\lambda}) = \frac{1}{2}(x_1^2 - x_2^2) - x_2 + \bar{\lambda}x_2 + \frac{c}{2}x_2^2 = \frac{1}{2}x_1^2 + \frac{1}{2}(c-1)x_2^2,$$

and we observe that \bar{x} is the unique unrestricted minimum of $L_c(\cdot, \bar{\lambda})$, for $c > \bar{c} = 1$.

6.3 Problems with Inequality Constraints

Consider the optimization problem (1.1) with inequality constraints

$$\begin{aligned} P : \text{Min } f(x) \\ \text{s.t. } g_i(x) \leq 0, \quad i = 1, 2, \dots, p, \end{aligned} \tag{6.32}$$

whose feasible set is

$$F := \{x \in \mathbb{R}^n : g_i(x) \leq 0, \quad i = 1, 2, \dots, p\}.$$

Throughout this section, we establish necessary conditions and sufficient conditions for a point $\bar{x} \in F$ to be a local minimum of P , i.e., for the existence of a neighborhood U of \bar{x} such that $f(\bar{x}) \leq f(x)$ for all $x \in F \cap U$.

The optimality conditions are a key tool for numerical methods. In fact, the verification of certain conditions of optimality is often used as stopping criterion in such methods. In this regard, the Karush–Kuhn–Tucker conditions (in brief, KKT conditions) play an important role in optimization as they become, under certain additional assumptions about the constraints of P , called *constraint qualifications*, necessary conditions of optimality, providing then a method for generating all candidates for local optima of P . In this section, we prove the nonconvex version of Theorem 4.46.

6.3.1 Karush–Kuhn–Tucker Optimality Conditions

Let us recall that $I(\bar{x})$ is the *set of active indices* at $\bar{x} \in F$, i.e.,

$$I(\bar{x}) := \{i \in \{1, 2, \dots, p\} : g_i(\bar{x}) = 0\},$$

whereas

$$A(\bar{x}) := \text{cone}\{\nabla g_i(\bar{x}) : i \in I(\bar{x})\},$$

is the *active cone* at \bar{x} .

Let us see that, under certain continuity assumption, in the search of local optima for P , we can eliminate the inactive constraints at \bar{x} . Formally, if $\bar{x} \in F$ is a local optimum of P , and g_i are continuous at \bar{x} , with $i \notin I(\bar{x})$, then the same point is a local optimum of the problem

$$\begin{aligned} P_{I(\bar{x})} : \text{Min } & f(x) \\ \text{s.t. } & g_i(x) \leq 0, i \in I(\bar{x}). \end{aligned}$$

Indeed, let U be a neighborhood of \bar{x} such that $f(\bar{x}) \leq f(x)$, for all $x \in F \cap U$, and let $V \subset \mathbb{R}^n$ be a neighborhood of \bar{x} such that $g_i(x) < 0$, for every $x \in V$ and $i \notin I(\bar{x})$ (the existence of such V is a consequence of the continuity of these functions at \bar{x}). Then, denoting by \bar{F} the feasible set of $P_{I(\bar{x})}$, we will have $f(\bar{x}) \leq f(x)$, for all $x \in \bar{F} \cap (V \cap U)$, since $\bar{F} \cap V \subset F$.

Moreover, it is obvious that \bar{x} is also a local minimum of the problem

$$\begin{aligned} \widehat{P}_{I(\bar{x})} : \text{Min } & f(x) \\ \text{s.t. } & g_i(x) = 0, i \in I(\bar{x}), \end{aligned}$$

since \bar{F} contains the feasible set of $\widehat{P}_{I(\bar{x})}$.

Thus, a possible first approach to the optimality conditions for the inequality constrained optimization problem P would consist in applying the Lagrange optimality conditions to the problem $\widehat{P}_{I(\bar{x})}$. More precisely, if $\bar{x} \in F$ is a local minimum of P , f is differentiable at \bar{x} , the functions g_i , $i \in I(\bar{x})$, are at least C^1 , the functions g_i , $i \notin I(\bar{x})$, are continuous at \bar{x} , and the gradients $\{\nabla g_i(\bar{x}) : i \in I(\bar{x})\}$ are linearly independent, then the Lagrange conditions (6.15) apply and yield the existence of scalars $\bar{\lambda}_i$, $i \in I(\bar{x})$, such that

$$\nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \bar{\lambda}_i \nabla g_i(\bar{x}) = 0_n. \quad (6.33)$$

(In the case $I(\bar{x}) = \emptyset$, we conclude $\nabla f(\bar{x}) = 0_n$.)

We shall see next that (6.33) can be refined, thanks to the generalized Farkas Lemma 4.15, by proving that $\bar{\lambda}_i \geq 0$, $i \in I(\bar{x})$, which gives rise to the so-called KKT conditions. We will also show that the condition of linear independence of the vectors $\{\nabla g_i(\bar{x}) : i \in I(\bar{x})\}$ is one of the announced constraint qualifications to be studied in the following section.

Optimization problems with inequality constraints were already considered by Fourier in 1798, in the context of analytical mechanics. In 1838, Ostrogradsky attempted to prove (6.33), for $n = 3$ and f being the gravitational potential, but his proof implicitly used Farkas lemma. Farkas observed its omission and published this result first in 1894. However, his proof contained a gap, which he finally fixed in 1901, after several unsuccessful attempts (see Prékopa [78] for more information about the beginnings of optimization theory).

As mentioned in Section 4.3, the systematic treatment of problems with inequality constraints was initiated by Karush in his master's thesis (1939) and Kuhn and Tucker [60]. These authors obtained, independently, the necessary conditions of optimality under certain constraint qualifications. Since the publication of Kuhn and Tucker's paper in 1951, various authors have devoted considerable effort to obtain such conditions under different assumptions of qualification (e.g., Cottle [26], Abadie [1], Mangasarian and Fromovitz [64], Guignard [49]).

The following cone plays a crucial role in our analysis. In the following definition, F is an arbitrary nonempty set in \mathbb{R}^n .

Definition 6.17 *Given $\bar{x} \in F$, we call tangent cone to F at \bar{x} to the set*

$$\mathcal{T}_{\bar{x}} := \left\{ d \in \mathbb{R}^n : d = \lim_{r \rightarrow \infty} \lambda_r(x_r - \bar{x}); \lambda_r > 0, x_r \in F \forall r, \text{ and } \lim_{r \rightarrow \infty} x_r = \bar{x} \right\}.$$

Proposition 6.18 *$\mathcal{T}_{\bar{x}}$ is a closed cone, not necessarily convex.*

Proof We can easily check that $\mathcal{T}_{\bar{x}}$ is a cone. Indeed, if $d \in \mathcal{T}_{\bar{x}}$, there exist $\lambda_r > 0$, $x_r \in F$, $r = 1, 2, \dots$, such that $d = \lim_{r \rightarrow \infty} \lambda_r(x_r - \bar{x})$, and then we have $\lambda d = \lim_{r \rightarrow \infty} \lambda \lambda_r(x_r - \bar{x}) \in \mathcal{T}_{\bar{x}}$, for all $\lambda > 0$. In addition, if $\lambda = 0$, then we write $\lambda d = 0_n = \lim_{r \rightarrow \infty} \lambda_r(\bar{x} - \bar{x}) \in \mathcal{T}_{\bar{x}}$. Hence, $\mathcal{T}_{\bar{x}}$ is cone. Example 6.31 shows that $\mathcal{T}_{\bar{x}}$ does not need to be convex.

Moreover, $\mathcal{T}_{\bar{x}}$ is closed. To prove this property, let $\{d_k\}$ be a sequence in $\mathcal{T}_{\bar{x}}$ converging to $d \in \mathbb{R}^n$. We can write

$$d_k = \lim_{r \rightarrow \infty} \lambda_{k,r}(x_{k,r} - \bar{x}), \quad k = 1, 2, \dots$$

Associated with each k , take r_k such that

$$\|d_k - \lambda_{k,r_k}(x_{k,r_k} - \bar{x})\| \leq \frac{1}{k} \quad \text{and} \quad \|x_{k,r_k} - \bar{x}\| \leq \frac{1}{k}.$$

Then,

$$\lim_{k \rightarrow \infty} \lambda_{k,r_k}(x_{k,r_k} - \bar{x}) = d \quad \text{and} \quad \lim_{k \rightarrow \infty} x_{k,r_k} = \bar{x},$$

which implies $d \in \mathcal{T}_{\bar{x}}$, and the proof is complete. \square

The following proposition expresses a first-order necessary optimality condition in terms of $\mathcal{T}_{\bar{x}}$. Therefore, the result is generally valid for an absolutely general feasible set F , not necessarily being described by explicit constraints.

Proposition 6.19 (Necessary optimality condition for an arbitrary feasible set F)
If $\bar{x} \in F \subset \mathbb{R}^n$ is a local minimum of the problem

$$\begin{aligned} P : \text{Min } & f(x) \\ \text{s.t. } & x \in F, \end{aligned}$$

and f is differentiable at \bar{x} , then

$$-\nabla f(\bar{x}) \in T_{\bar{x}}^{\circ}.$$

Proof We have to prove that $\nabla f(\bar{x})^T d \geq 0$ for every $d \in T_{\bar{x}}$. By definition, $d = \lim_{r \rightarrow \infty} \lambda_r(x_r - \bar{x})$ for certain $\lambda_r > 0$, $x_r \in F$ for all r , and $\lim_{r \rightarrow \infty} x_r = \bar{x}$. We consider the nontrivial case $d \neq 0_n$, which allows to suppose without loss of generality that $x_r - \bar{x} \neq 0$ for all r . The differentiability of f at \bar{x} yields

$$f(x_r) = f(\bar{x}) + \nabla f(\bar{x})^T (x_r - \bar{x}) + o(\|x_r - \bar{x}\|). \quad (6.34)$$

Since \bar{x} is a local minimum of P and $x_r \in F$, it must be $f(x_r) \geq f(\bar{x})$ for r large enough, e.g., for $r \geq r_0$. By (6.34), we get $\nabla f(\bar{x})^T (x_r - \bar{x}) + o(\|x_r - \bar{x}\|) \geq 0$, for $r \geq r_0$. Then,

$$\nabla f(\bar{x})^T d = \lim_{r \rightarrow \infty} \left\{ \lambda_r \nabla f(\bar{x})^T (x_r - \bar{x}) + \lambda_r \|x_r - \bar{x}\| \frac{o(\|x_r - \bar{x}\|)}{\|x_r - \bar{x}\|} \right\} \geq 0,$$

because $\lim_{r \rightarrow \infty} \lambda_r \|x_r - \bar{x}\| = \|d\|$. □

The condition in the latter proposition, although it does not directly lead to a practical method for solving the problem P (as $T_{\bar{x}}^{\circ}$ is difficult to characterize), will be of great theoretical utility in the rest of this chapter.

Definition 6.20 The point $\bar{x} \in F$ is said to be a KKT point of the problem P with inequality constraints (6.32) if there exist scalars $\bar{\lambda}_i \geq 0$, $i \in I(\bar{x})$, such that

$$-\nabla f(\bar{x}) = \sum_{i \in I(\bar{x})} \bar{\lambda}_i \nabla g_i(\bar{x}),$$

or, equivalently, if

$$-\nabla f(\bar{x}) \in A(\bar{x}). \quad (6.35)$$

In Subsection 4.3.1, where optimality conditions for problems with linear inequality constraints were established, (6.35) was called *KKT condition*. Clearly, we have that (6.35) can be equivalently expressed as

$$-\nabla f(\bar{x}) = \sum_{i=1}^p \bar{\lambda}_i \nabla g_i(\bar{x}), \quad \bar{\lambda}_i g_i(\bar{x}) = 0, \quad \bar{\lambda}_i \geq 0, \quad i = 1, 2, \dots, p.$$

Conditions $\bar{\lambda}_i g_i(\bar{x}) = 0$, $i = 1, 2, \dots, p$, were referred to as *complementarity conditions* in Theorem 4.46.

Now, we introduce a new cone, $G_{\bar{x}}$, which allows us to connect the optimality condition given in Proposition 6.19 with the KKT conditions:

$$G_{\bar{x}} := \{d \in \mathbb{R}^n : \nabla g_i(\bar{x})^T d \leq 0, i \in I(\bar{x})\}.$$

Observe that $\bar{x} \in F$ is a KKT point of P if and only if

$$-\nabla f(\bar{x}) \in G_{\bar{x}}^\circ,$$

since, by Farkas Lemma 4.15,

$$A(\bar{x}) = \text{cone}\{\nabla g_i(\bar{x}), i \in I(\bar{x})\} = \{\nabla g_i(\bar{x}), i \in I(\bar{x})\}^{\circ\circ} = G_{\bar{x}}^\circ.$$

The following example illustrates the necessary optimality condition established in Proposition 6.19, while showing a situation where the KKT conditions fail.

Example 6.21 ([60]) Consider the optimization problem in \mathbb{R}^2 :

$$\begin{aligned} P : \text{Min } & x_1 \\ \text{s.t. } & x_2 - x_1^3 \leq 0, \\ & -x_2 \leq 0. \end{aligned}$$

We can easily verify that at $\bar{x} = (0, 0)^T$ the tangent cone is $T_{\bar{x}} = \text{cone}\{(1, 0)^T\}$, whereas $G_{\bar{x}} = \text{span}\{(1, 0)^T\}$. Hence, $-\nabla f(\bar{x}) = (-1, 0)^T \in T_{\bar{x}}^\circ$, but $-\nabla f(\bar{x}) \notin G_{\bar{x}}^\circ$, and $\bar{x} = (0, 0)^T$ is not a KKT point. Moreover, it is easy to check that \bar{x} is a local optimum (actually, it is a global minimum because every feasible point satisfies $x_1^3 \geq x_2 \geq 0$, i.e., $x_1 \geq 0$).

Taking into account the last observation, it is obvious that if $T_{\bar{x}}^\circ = G_{\bar{x}}^\circ$ the KKT conditions turn out to be necessary for \bar{x} to be a local minimum. Additionally, the equality $T_{\bar{x}}^\circ = G_{\bar{x}}^\circ$ is equivalent to $\text{cl}(\text{cone}T_{\bar{x}}) = G_{\bar{x}}$. Indeed, if $T_{\bar{x}}^\circ = G_{\bar{x}}^\circ$, and we apply Farkas Lemma 4.15 and Proposition 4.17(iii), one has $\text{cl}(\text{cone}T_{\bar{x}}) = T_{\bar{x}}^{\circ\circ} = G_{\bar{x}}^{\circ\circ} = G_{\bar{x}}$. Conversely, if $\text{cl}(\text{cone}T_{\bar{x}}) = G_{\bar{x}}$, then $T_{\bar{x}}^\circ = (\text{cl}(\text{cone}T_{\bar{x}}))^\circ = G_{\bar{x}}^\circ$, just by applying Proposition 4.17(ii), and we are done. Actually, we have proved the following result:

Theorem 6.22 (KKT necessary optimality conditions) *Let $\bar{x} \in F$ be a local minimum of the problem with inequality constraints (6.32). Assume that the functions f and g_i , $i \in I(\bar{x})$, are differentiable at \bar{x} and that $\text{cl}(\text{cone}T_{\bar{x}}) = G_{\bar{x}}$. Then, \bar{x} is a KKT point.*

In this way, the equality $\text{cl}(\text{cone}T_{\bar{x}}) = G_{\bar{x}}$ constitutes a hypothesis of constraint qualification which is known as *Guignard constraint qualification (GCQ, in brief)*. This hypothesis is the weakest among all the constraint qualifications considered in the next subsection.

6.3.2 Other Constraint Qualifications*

Now, we proceed by introducing some more constraint qualifications. To this aim, let us consider the following pair of sets associated with each $\bar{x} \in F$: the strict polar of the active cone

$$\tilde{G}_{\bar{x}} := \{d \in \mathbb{R}^n : \nabla g_i(\bar{x})^T d < 0, i \in I(\bar{x})\},$$

and the cone of tangential directions

$$T_{\bar{x}} := \left\{ d \in \mathbb{R}^n \middle| \begin{array}{l} \exists \varepsilon > 0, \exists \alpha : [0, \varepsilon] \rightarrow F \text{ differentiable in } [0, \varepsilon], \\ \text{with } \alpha(0) = \bar{x} \text{ and } \alpha'(0) = d \end{array} \right\}.$$

Obviously, $\tilde{G}_{\bar{x}}$ is an open and convex set, and $\tilde{G}_{\bar{x}} \cup \{0_n\}$ is a cone. Observe also that $T_{\bar{x}}$, like $\mathcal{T}_{\bar{x}}$, does not rely on an explicit representation of F ; i.e., F can be an arbitrary set in \mathbb{R}^n .

Proposition 6.23 $T_{\bar{x}}$ is a cone, not necessarily closed.

Proof If $d \in T_{\bar{x}}$, there exists a function $\alpha : [0, \varepsilon] \rightarrow F$, for a certain $\varepsilon > 0$, differentiable in $[0, \varepsilon]$ and such that $\alpha'(0) = d$ and $\alpha(0) = \bar{x}$. If $\lambda > 0$, then the function: $[0, \frac{\varepsilon}{\lambda}] \rightarrow F$, given by $\beta(t) = \alpha(\lambda t)$, verifies $\beta(0) = \alpha(0) = \bar{x}$, and $\beta'(0) = \lambda \alpha'(0) = \lambda d \in T_{\bar{x}}$. If $\lambda = 0$, it is enough to consider $\alpha : [0, \varepsilon] \rightarrow F$ such that $\alpha(t) = \bar{x}$, for all $t \in [0, \varepsilon]$ (with ε arbitrary); then, $\lambda d = 0_n \in T_{\bar{x}}$.

However, $T_{\bar{x}}$ is not closed in general. This is the case for

$$F = \{x \in \mathbb{R}^2 : g_1(x_1, x_2) = 0, x_1 - x_2^2 \geq 0\},$$

where

$$g_1(x_1, x_2) = \begin{cases} x_1^2 \sin\left(\pi \frac{x_2}{x_1}\right), & \text{if } x_1 \neq 0, \\ 0, & \text{if } x_1 = 0, \end{cases}$$

see Fig. 6.6.

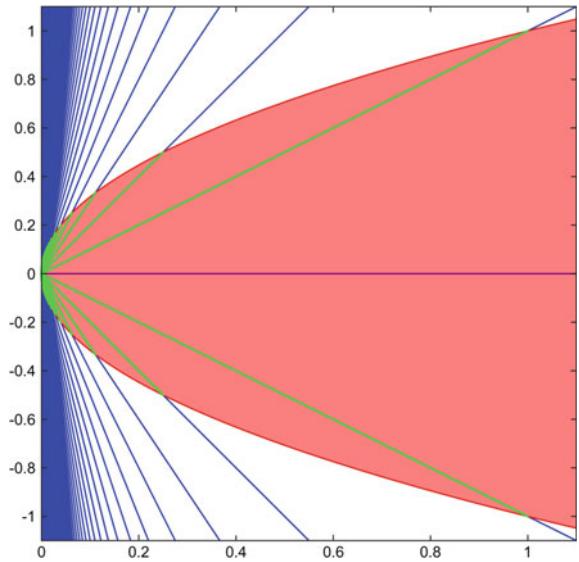
The reader can verify that

$$F = \{0_2\} \cup \bigcup_{r \in \mathbb{Z}} \left\{ x \in \mathbb{R}^2 : x_2 = rx_1, 0 \leq x_1 \leq \frac{1}{r^2} \right\}.$$

Therefore, if $\bar{x} = 0_2$ we have $d_r := r\left(\frac{1}{r^2}, \frac{1}{r}\right)^T = \left(\frac{1}{r}, 1\right)^T \in T_{\bar{x}}$, $r = 1, 2, \dots$, but $\lim_{r \rightarrow \infty} d_r = \lim_{r \rightarrow \infty} \left(\frac{1}{r}, 1\right)^T = (0, 1)^T \notin T_{\bar{x}}$. \square

Next, we study the inclusion relations among these sets. We shall assume from now on that $\bar{x} \in F$ and that the functions g_i , $i \in I(\bar{x})$, are differentiable at \bar{x} , and the functions g_i , $i \notin I(\bar{x})$, are continuous at this point.

Fig. 6.6 Feasible set (the union of the green segments) for which $T_{\bar{x}}$ is not closed



Proposition 6.24 *The inclusion $\tilde{G}_{\bar{x}} \subset T_{\bar{x}}$ holds.*

Proof Take $d \in \tilde{G}_{\bar{x}}$ and define $\alpha(t) := \bar{x} + td$, with $t \geq 0$. If $i \in I(\bar{x})$,

$$g_i(\bar{x} + td) = g_i(\bar{x}) + t \nabla g_i(\bar{x})^T d + o(t) = t \left(\nabla g_i(\bar{x})^T d + \frac{o(t)}{t} \right),$$

and taking t small enough, we get

$$\nabla g_i(\bar{x})^T d + \frac{o(t)}{t} < 0.$$

As $I(\bar{x})$ is a finite set, there exists $\varepsilon > 0$ such that

$$g_i(\bar{x} + td) < 0, \quad \text{for all } t \in [0, \varepsilon] \text{ and } i \in I(\bar{x}).$$

Since the functions g_i , $i \notin I(\bar{x})$, are continuous at \bar{x} , if ε is small enough, we also have

$$g_i(\bar{x} + td) < 0, \quad \text{for all } t \in [0, \varepsilon] \text{ and } i \notin I(\bar{x}).$$

Hence, $\alpha(t) \in F$ for $t \in [0, \varepsilon]$. At the same time, $\alpha(0) = \bar{x}$ and $\alpha'(0) = d$, and we conclude that $d \in T_{\bar{x}}$. \square

Theorem 6.25 (Basic inclusions) *The following inclusion relations hold:*

$$\text{cl}(\tilde{G}_{\bar{x}}) \subset T_{\bar{x}} \subset \mathcal{T}_{\bar{x}} \subset \text{cl}(\text{cone} \mathcal{T}_{\bar{x}}) \subset G_{\bar{x}}.$$

Proof • The proof of the inclusion $\text{cl}(\tilde{G}_{\bar{x}}) \subset T_{\bar{x}}$ is out of the scope of this book (it can be found in [76, Theorem 3]). Observe that this is not a direct consequence of the inclusion $\tilde{G}_{\bar{x}} \subset T_{\bar{x}}$, as we showed that $T_{\bar{x}}$ is not closed in general.

- The inclusion $T_{\bar{x}} \subset \tilde{T}_{\bar{x}}$ is obvious since $d \in T_{\bar{x}}$ entails $d = \lim_{t \searrow 0} \frac{\alpha(t) - \alpha(0)}{t}$ for a certain function $\alpha : [0, \varepsilon] \rightarrow F$ (with $\varepsilon > 0$) and, in particular, one has

$$d = \lim_{r \rightarrow \infty} \frac{r}{\varepsilon} \left(\alpha\left(\frac{\varepsilon}{r}\right) - \alpha(0) \right) \in \tilde{T}_{\bar{x}}.$$

- Let us prove now that $\text{cl}(\text{cone } \tilde{T}_{\bar{x}}) \subset G_{\bar{x}}$. Since $G_{\bar{x}}$ is a closed convex cone, it is enough to prove that $\tilde{T}_{\bar{x}} \subset G_{\bar{x}}$. If $d \in \tilde{T}_{\bar{x}}$, we can write $d = \lim_{r \rightarrow \infty} \lambda_r(x_r - \bar{x})$ with $\lambda_r > 0$, $x_r \in F$ for all r , and $\lim_{r \rightarrow \infty} x_r = \bar{x}$. The differentiability of g_i at \bar{x} , for $i \in I(\bar{x})$, yields

$$g_i(x_r) = g_i(\bar{x}) + \nabla g_i(\bar{x})^T(x_r - \bar{x}) + o(\|x_r - \bar{x}\|), \quad r = 1, 2, \dots \quad (6.36)$$

Since $g_i(x_r) \leq 0$, $r = 1, 2, \dots$, multiplying both members in (6.36) by λ_r and letting $r \rightarrow +\infty$, one gets

$$\nabla g_i(\bar{x})^T d = \lim_{r \rightarrow \infty} \left\{ \lambda_r \nabla g_i(\bar{x})^T(x_r - \bar{x}) + \|\lambda_r(x_r - \bar{x})\| \frac{o(\|x_r - \bar{x}\|)}{\|x_r - \bar{x}\|} \right\} \leq 0,$$

leading to conclude that $d \in G_{\bar{x}}$. \square

The inclusion relations between the sets established above lead to the following hypotheses defining constraint qualifications at \bar{x} and determine the implications shown below the following table in (6.37):

Constr. qualification	In brief	Hypothesis
<i>Mangasarian–Fromovitz</i> (also, <i>Cottle</i>)	MFCQ	$\text{cl}(\tilde{G}_{\bar{x}}) = G_{\bar{x}}$ $(\iff \tilde{G}_{\bar{x}} \neq \emptyset)$
<i>Kuhn–Tucker</i>	KTCQ	$T_{\bar{x}} = G_{\bar{x}}$
<i>Abadie</i>	ACQ	$\tilde{T}_{\bar{x}} = G_{\bar{x}}$
<i>Guignard</i>	GCQ	$\text{cl}(\text{cone } \tilde{T}_{\bar{x}}) = G_{\bar{x}}$

$$\text{MFCQ} \Rightarrow \text{KTCQ} \Rightarrow \text{ACQ} \Rightarrow \text{GCQ}. \quad (6.37)$$

The equivalence

$$\text{cl}(\tilde{G}_{\bar{x}}) = G_{\bar{x}} \iff \tilde{G}_{\bar{x}} \neq \emptyset,$$

is established in Exercise 6.8.

Next, we introduce new hypotheses of constraint qualification, which constitute sufficient conditions for any of the aforementioned conditions and that, in certain situations, may be more operational. One is based on the following theorem. Note also that the statement of this theorem provides a characterization of the condition $\tilde{G}_{\bar{x}} \neq \emptyset$ (MFCQ).

Theorem 6.26 (Gordan theorem) *The system of linear strict inequalities, in \mathbb{R}^n , $\{a_i^T x < 0; i = 1, 2, \dots, p\}$ has no solution if and only if there exist nonnegative scalars $\lambda_1, \dots, \lambda_p$, not all of them equal to zero, such that $\sum_{i=1}^p \lambda_i a_i = 0_n$.*

Proof Assume first that there exists some $x \in \mathbb{R}^n$ such that $a_i^T x < 0$ for all $i = 1, \dots, p$. If there exist nonnegative scalars $\lambda_1, \dots, \lambda_p$, not all of them zero, such that $\sum_{i=1}^p \lambda_i a_i = 0_n$, we obtain a contradiction with the fact that $\sum_{i=1}^p \lambda_i a_i^T x < 0$.

Suppose now that the system $\{a_i^T x < 0, i = 1, \dots, p\}$ has no solution. Hence, the system $\{a_i^T x + s \leq 0, i = 1, \dots, p; s > 0\}$ cannot have a solution either; that is,

$$\forall x \in \mathbb{R}^n : \left[\begin{pmatrix} a_i \\ 1 \end{pmatrix}^T \begin{pmatrix} x \\ s \end{pmatrix} \leq 0, \quad i = 1, \dots, p \right] \Rightarrow \left[\begin{pmatrix} 0_n \\ 1 \end{pmatrix}^T \begin{pmatrix} x \\ s \end{pmatrix} \leq 0 \right].$$

By Corollary 4.16, there are nonnegative scalars $\lambda_1, \dots, \lambda_p$ such that

$$\sum_{i=1}^p \lambda_i \begin{pmatrix} a_i \\ 1 \end{pmatrix} = \begin{pmatrix} 0_n \\ 1 \end{pmatrix},$$

from where we deduce that $\sum_{i=1}^p \lambda_i a_i = 0_n$, with $\lambda_j > 0$ for some $j \in \{1, \dots, p\}$. \square

Proposition 6.27 *If any of the following properties is fulfilled, we have $\tilde{G}_{\bar{x}} \neq \emptyset$:*

- (i) *The vectors $\{\nabla g_i(\bar{x}), i \in I(\bar{x})\}$ are linearly independent.*
- (ii) *The functions $g_i, i \in I(\bar{x})$, are convex, and there is a point \hat{x} such that $g_i(\hat{x}) < 0$, $i \in I(\bar{x})$.*

Proof The implication (i) \Rightarrow $\{\tilde{G}_{\bar{x}} \neq \emptyset\}$ is straightforward from Gordan Theorem 6.26.

Let us prove the second implication (ii) \Rightarrow $\{\tilde{G}_{\bar{x}} \neq \emptyset\}$. Let $\hat{x} \in \mathbb{R}^n$ be such that $g_i(\hat{x}) < 0$, $i \in I(\bar{x})$. Since we are assuming that the functions $g_i, i \in I(\bar{x})$, are convex and differentiable at \bar{x} , we have by Proposition 2.33 that

$$g_i(\bar{x}) + \nabla g_i(\bar{x})^T (\hat{x} - \bar{x}) \leq g_i(\hat{x}), \quad \text{for all } x \in \mathbb{R}^n.$$

In particular, $\nabla g_i(\bar{x})^T (\hat{x} - \bar{x}) = g_i(\bar{x}) + \nabla g_i(\bar{x})^T (\hat{x} - \bar{x}) \leq g_i(\hat{x}) < 0$, for $i \in I(\bar{x})$; i.e., $\hat{x} - \bar{x} \in \tilde{G}_{\bar{x}}$. \square

Conditions (i) and (ii) in Proposition 6.27 provide the following two additional constraint qualifications:

- *Linear independence constraint qualification (LICQ):* The vectors $\{\nabla g_i(\bar{x}), i \in I(\bar{x})\}$ are linearly independent.
- *Slater constraint qualification (SCQ):* The functions $g_i, i \in I(\bar{x})$, are convex, and there is a point \hat{x} such that $g_i(\hat{x}) < 0$, $i \in I(\bar{x})$.

The last proposition completes (6.37) with two new implications:

$$\begin{array}{c} \text{LICQ} \Rightarrow \text{MFCQ} \\ \uparrow \\ \text{SCQ} \end{array} \quad (6.38)$$

Theorem 6.28 (Sufficient conditions for KKT points) *Let $\bar{x} \in F$ be a local minimum of the problem with inequality constraints (6.32). Assume that the functions f and g_i , $i \in I(\bar{x})$, are differentiable at \bar{x} and that g_i , $i \notin I(\bar{x})$, are continuous at \bar{x} . If any of the qualifications introduced in the previous paragraphs are fulfilled, then \bar{x} is a KKT point.*

To illustrate the above theorem and show that none of the converse implications in (6.37) and (6.38) hold, we present the following examples.

Example 6.29 (MFCQ holds, and LICQ and SCQ fail) Consider the optimization problem in \mathbb{R}^2 :

$$\begin{aligned} P : \text{Min } & x_1 \\ \text{s.t. } & x_2 - x_1^3 \leq 0, \\ & -x_1 \leq 0, \\ & -x_1 + x_2 \leq 0. \end{aligned}$$

At $\bar{x} = 0_2$, the active indices are $I(\bar{x}) = \{1, 2, 3\}$, and $\nabla g_1(\bar{x}) = (0, 1)^T$, $\nabla g_2(\bar{x}) = (-1, 0)^T$, $\nabla g_3(\bar{x}) = (-1, 1)^T$, which are linearly dependent. Hence, LICQ fails. Moreover, SCQ also fails as g_1 is not convex. However,

$$\tilde{G}_{\bar{x}} = \{d \in \mathbb{R}^2 : d_2 < 0, -d_1 < 0, -d_1 + d_2 < 0\} \neq \emptyset,$$

and so, MFCQ holds. The feasible set of P and $\tilde{G}_{\bar{x}}$ are represented in Fig. 6.7. Observe that \bar{x} is a local optimum of P and also a KKT point.

Example 6.30 (KTCQ holds, and MFCQ fails) Let us consider the problem in \mathbb{R}^2 :

$$\begin{aligned} P : \text{Min } & x_1 \\ \text{s.t. } & x_2 - x_1^3 \leq 0, \\ & -x_1 \leq 0, \\ & -x_2 \leq 0. \end{aligned}$$

Take $\bar{x} = 0_2$. It is obvious that $\tilde{G}_{\bar{x}} = \emptyset$ because $\nabla g_1(\bar{x}) = (0, 1)^T$ and $\nabla g_3(\bar{x}) = (0, -1)^T$; i.e., MFCQ fails. On the other hand, $G_{\bar{x}} = \text{cone}\{(1, 0)^T\} \subset T_{\bar{x}}$, since $\bar{x} + t(1, 0)^T \in F$ for $t \in [0, +\infty[$; thus, KTCQ holds. The point $\bar{x} = 0_2$ is a local

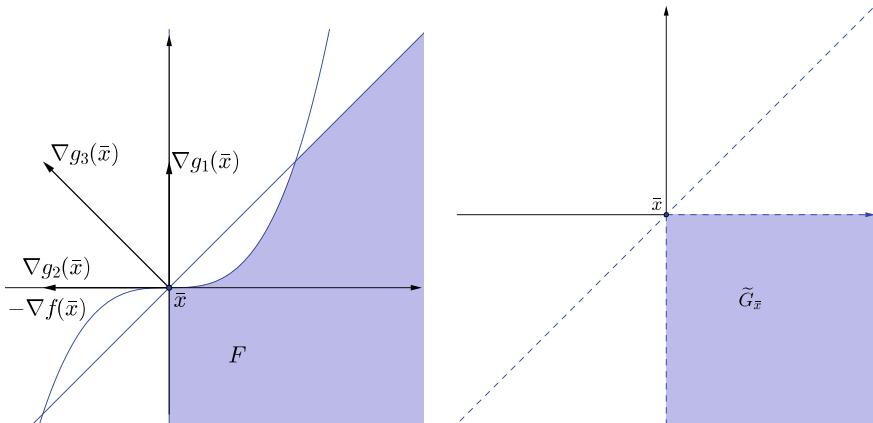


Fig. 6.7 Feasible set and $\tilde{G}_{\bar{x}}$ for Example 6.29

minimum of P and also a KKT point. Observe that the only difference between this example and Example 6.21 is that now we have added the redundant constraint $-x_1 \leq 0$.

Example 6.31 (GCQ holds, and ACQ fails) Consider the problem in \mathbb{R}^2 :

$$\begin{aligned} P : \text{Min } & x_1 \\ \text{s.t. } & x_1 x_2 \leq 0, \\ & -x_1 x_2 \leq 0, \\ & -x_1 \leq 0, \\ & -x_2 \leq 0. \end{aligned}$$

It is easy to verify that

$$F = \{x \in \mathbb{R}^2 : x_1 \geq 0, x_2 = 0\} \cup \{x \in \mathbb{R}^2 : x_1 = 0, x_2 \geq 0\}.$$

Then, we observe that, at $\bar{x} = 0_2$, $\mathcal{T}_{\bar{x}} = \text{cone}\{(1, 0)^T\} \cup \text{cone}\{(0, 1)^T\}$, whereas $G_{\bar{x}} = \text{cone}\{(1, 0)^T, (0, 1)^T\}$. Hence, ACQ fails but $\text{cl}(\text{cone}\mathcal{T}_{\bar{x}}) = G_{\bar{x}}$; i.e., GCQ holds.

In the above examples, we were investigating whether any constraint qualification was verified and if the KKT conditions hold or not at a given point \bar{x} . However, when we are solving an optimization problem, we have to search all candidates for being local minima by analyzing all possible choices of active indices. Thus, according to the results presented in this section, we will consider as candidates those points that simultaneously satisfy some constraint qualifications and the KKT conditions. We illustrate this procedure with the following example.

Example 6.32 Consider the problem in \mathbb{R}^2 :

$$\begin{aligned} P : \text{Min } & x_2 \\ \text{s.t. } & -x_1^2 - x_2^2 + 1 \leq 0, \\ & (x_1 - 1)^2 + x_2^2 - 1 \leq 0, \\ & -2(x_1 - \frac{1}{2})^3 + x_2^2 - \frac{3}{4} \leq 0. \end{aligned}$$

We shall analyze the different sets of active indices (since there are three constraints, we need to analyze $2^3 = 8$ cases).

- (1) $I(x) = \emptyset$. No point verifies $\nabla f(x) = 0_2$.
- (2) $I(x) = \{1\}$. LICQ holds and the candidates to be local minima are all the KKT points of this type. They must satisfy $(0, -1) = \lambda_1(-2x_1, -2x_2)$. The unique solution of this equation with $\lambda_1 \geq 0$ is $(x_1, x_2, \lambda_1) = (0, 1, 1/2)$, which fails to provide a feasible point. Therefore, no candidate is envisaged in this case.
- (3) $I(x) = \{2\}$. Now, SCQ holds (g_2 is convex and, for instance, $g_2(1, 0) = -1 < 0$). In this case, the KKT conditions give rise to $(x_1, x_2, \lambda_2) = (1, -1, 1/2)$, which does not correspond with this case because the third constraint is also active at $(1, -1)^T$.
- (4) $I(x) = \{3\}$. LICQ is fulfilled as the unique solution of $\nabla g_3(x) = 0_2$ is $x = (1/2, 0)^T$, which is not feasible. The system $-\nabla f(x) = \lambda_3 \nabla g_3(x)$ yields two solutions $(x_1, x_2, \lambda_3) = \left(\frac{1}{2}, \frac{\sqrt{3}}{2}, -\frac{1}{\sqrt{3}}\right)$ and $(x_1, x_2, \lambda_3) = \left(\frac{1}{2}, -\frac{\sqrt{3}}{2}, \frac{1}{\sqrt{3}}\right)$. The first triple is not a KKT point since $\lambda_3 < 0$. The second point $\left(\frac{1}{2}, \frac{\sqrt{3}}{2}\right)^T$ also activates the other two constraints and so, it does not correspond with this case. Similar situations arise in the cases (5) $I(x) = \{1, 2\}$ and (6) $I(x) = \{1, 3\}$.
- (7) $I(x) = \{2, 3\}$. LICQ is satisfied, and the KKT conditions provide a unique candidate in this case: $(x_1, x_2, \lambda_2, \lambda_3) = (1, -1, 1/2, 0)$.
- (8) $I(x) = \{1, 2, 3\}$. Both SCQ and LICQ fail now, but we can easily verify that MFCQ holds. The unique KKT point in this case is $x = \left(\frac{1}{2}, -\frac{\sqrt{3}}{2}\right)^T$.

We finally obtained two candidates for local optima: $(1, -1)^T$ and $\left(1/2, -\sqrt{3}/2\right)^T$. In Fig. 6.8, we may appreciate that $\left(1/2, -\sqrt{3}/2\right)^T$ is not a local minimum. More formally, for the sequence of feasible points

$$x_r := \left(\frac{1}{2} + \frac{1}{r}, -\sqrt{\frac{3}{4} + \frac{2}{r^3}}\right)^T, \quad r = 2, 3, \dots,$$

converging to $\bar{x} = \left(\frac{1}{2}, -\sqrt{\frac{3}{4}}\right)^T$, we observe that $f(x_r) < f(\bar{x})$, for all $r \geq 2$.

In fact, $x = (1, -1)^T$ is a global minimum of P . In this particular case, we do not need to make any additional discussion, since F is compact, so Weierstrass theorem

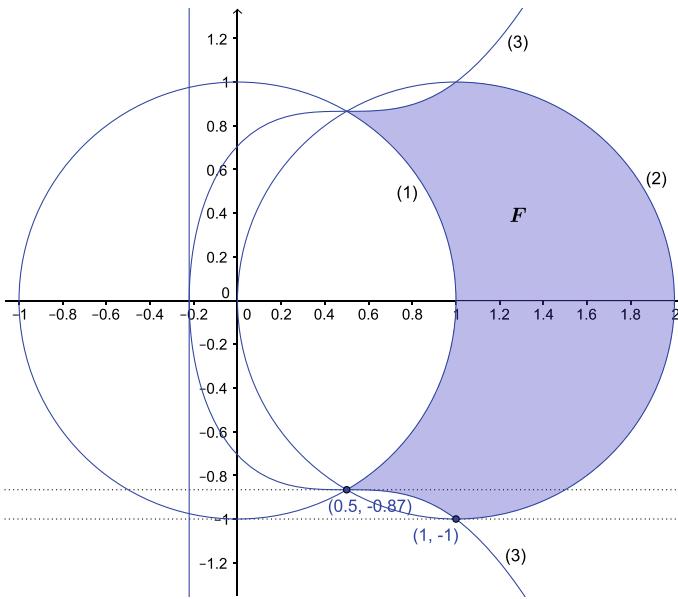


Fig. 6.8 Illustrating Example 6.32

applies to conclude the existence of a global minimum of P . Since $(1, -1)^T$ is the only candidate, it must be a global optimum of P .

In the above example, we have found that the KKT necessary optimality conditions are not, however, sufficient. The following theorem, consequence of Theorem 4.46 and Remark 4.47, shows how under the convexity assumption, the KKT conditions are *sufficient*, not only for local optimality, but directly to global optimality.

Theorem 6.33 (Global optimality conditions) *If \bar{x} is a KKT point for P and we assume that the functions f and g_i , $i \in I(\bar{x})$, are convex and differentiable at \bar{x} , then \bar{x} is a global minimum of P .*

6.3.3 Fritz John Optimality Conditions*

Complementing the results in the chapter, we present a new necessary optimality condition in the line of the KKT conditions, although weaker. As a counterpart, it requires no constraint qualification.

Theorem 6.34 (Fritz John optimality conditions) *Let \bar{x} be a local minimum of the problem P in (6.32), and suppose that f , g_i , $i \in I(\bar{x})$, are differentiable at \bar{x} , and g_i , $i \notin I(\bar{x})$, are continuous at \bar{x} . Then, there exist scalars $\bar{\lambda}_0, \bar{\lambda}_i \geq 0$, $i \in I(\bar{x})$, not all null, such that*

$$\bar{\lambda}_0 \nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \bar{\lambda}_i \nabla g_i(\bar{x}) = 0_n.$$

Proof Under the current assumptions, Proposition 6.19 states that $-\nabla f(\bar{x}) \in T_{\bar{x}}^\circ$. On the other hand, Theorem 6.25 shows that $\tilde{G}_{\bar{x}} \subset T_{\bar{x}} \subset T_{\bar{x}}^\circ$, which entails $T_{\bar{x}}^\circ \subset (\tilde{G}_{\bar{x}})^\circ$; so, we have that

$$-\nabla f(\bar{x}) \in (\tilde{G}_{\bar{x}})^\circ.$$

In other words, $\nabla f(\bar{x})^T d \geq 0$ for all $d \in \mathbb{R}^n$ verifying $\nabla g_i(\bar{x})^T d < 0$ for all $i \in I(\bar{x})$; that is, the system

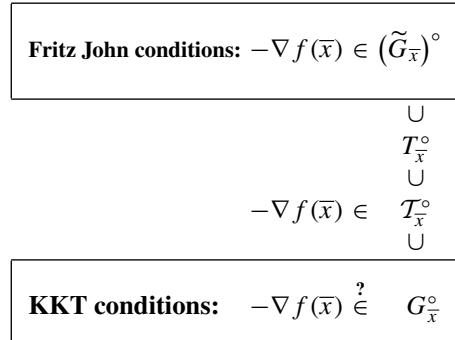
$$\{\nabla f(\bar{x})^T d < 0; \nabla g_i(\bar{x})^T d < 0, i \in I(\bar{x})\}$$

has no solution $d \in \mathbb{R}^n$. Then, by Gordan Theorem 6.26, there exist $\bar{\lambda}_0, \bar{\lambda}_i \geq 0$, $i \in I(\bar{x})$, not all zero, such that $\bar{\lambda}_0 \nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \bar{\lambda}_i \nabla g_i(\bar{x}) = 0_n$. \square

The following diagram is intended to show the relationships between the conditions of Fritz John and the other necessary optimality conditions studied in this chapter. Once again, we are assuming that \bar{x} is a local optimum of problem P and that $f, g_i, i \in I(\bar{x})$, are differentiable at \bar{x} , whereas $g_i, i \notin I(\bar{x})$, are continuous at \bar{x} . Under these assumptions, the diagram is a direct consequence of the following inclusions

$$\text{cl}(\tilde{G}_{\bar{x}}) \subset T_{\bar{x}} \subset T_{\bar{x}}^\circ \subset \text{cl}(\text{cone } T_{\bar{x}}) \subset G_{\bar{x}},$$

which were established in Theorem 6.25.



6.4 Problems with Equality and Inequality Constraints

In this section, we deal with the optimization problems introduced in (1.1), i.e., with optimization problems of the form

$$\begin{aligned} P : \text{Min } & f(x) \\ \text{s.t. } & h_j(x) = 0, j = 1, \dots, m, \\ & g_i(x) \leq 0, i = 1, \dots, p. \end{aligned} \tag{6.39}$$

Here, the set C is the whole space \mathbb{R}^n .

We might think of replacing each of the equalities $h_j(x) = 0$ by two inequalities $h_j(x) \leq 0$ and $-h_j(x) \leq 0$, and once P is adapted to the format discussed in the previous subsection, then apply the results obtained there. This procedure is possible, and even more, since the feasible region of P is not affected by this new representation, given $\bar{x} \in F$, the sets $T_{\bar{x}}$ and $\mathcal{T}_{\bar{x}}$ are defined in the same way, they have the same properties and the same relationship between them (i.e., $T_{\bar{x}} \subset \mathcal{T}_{\bar{x}}$).

However, from this representation in terms of inequalities, if we try to adapt the set $\tilde{G}_{\bar{x}}$ to the new setting, we find that it is necessarily empty. Thus, if we want to introduce qualifications in line with the hypothesis $\tilde{G}_{\bar{x}} \neq \emptyset$, equality constraints must be treated as equalities. To this end, we consider the following sets:

$$\begin{aligned} \tilde{G}_{\bar{x}} &:= \{d \in \mathbb{R}^n : \nabla g_i(\bar{x})^T d < 0, i \in I(\bar{x})\}, \\ G_{\bar{x}} &:= \{d \in \mathbb{R}^n : \nabla g_i(\bar{x})^T d \leq 0, i \in I(\bar{x})\}, \\ H_{\bar{x}} &:= \{d \in \mathbb{R}^n : \nabla h_j(\bar{x})^T d = 0, j = 1, 2, \dots, m\}, \end{aligned}$$

where $I(\bar{x}) := \{i \in \{1, \dots, p\} : g_i(\bar{x}) = 0\}$.

We begin by noting that Theorem 6.22, which establishes the KKT conditions as necessary optimality conditions under the Guignard constraint qualification (which at that time was formulated as $\text{cl}(\text{cone}\mathcal{T}_{\bar{x}}) = G_{\bar{x}}$), can easily be adapted to the new context. Following the steps in the proof of that theorem, and decomposing each equality $h_j(x) = 0$ into two inequalities $h_j(x) \leq 0$ and $-h_j(x) \leq 0$, the new statement would read as follows:

Theorem 6.35 (KKT points) *Let $\bar{x} \in F$ be a local minimum of the problem (6.39). Suppose that the functions f, g_i , with $i \in I(\bar{x})$, and h_j , $j = 1, \dots, m$, are differentiable at \bar{x} and that the equality $\text{cl}(\text{cone}\mathcal{T}_{\bar{x}}) = G_{\bar{x}} \cap H_{\bar{x}}$ holds true. Then, there are scalars $\bar{\lambda}_i \geq 0$, $i \in I(\bar{x})$, $\bar{\mu}_j \in \mathbb{R}$, $j = 1, 2, \dots, m$, such that*

$$-\nabla f(\bar{x}) = \sum_{i \in I(\bar{x})} \bar{\lambda}_i \nabla g_i(\bar{x}) + \sum_{j=1}^m \bar{\mu}_j \nabla h_j(\bar{x}).$$

We then say that \bar{x} is a KKT point of the problem (6.39).

The following result gives inclusion relations among the sets used in this section, which will lead to new constraint qualifications. Note that this new proposition adapts to the context of the problem (6.39) the conditions of Theorem 6.25.

Theorem 6.36 (Advanced inclusions) *Let \bar{x} be a feasible point of problem (6.39). Suppose that the functions g_i , with $i \in I(\bar{x})$, and h_j , $j = 1, \dots, m$, are differentiable at \bar{x} . Then, the following statements are true:*

- (i) $T_{\bar{x}} \subset \mathcal{T}_{\bar{x}} \subset \text{cl}(\text{cone}\mathcal{T}_{\bar{x}}) \subset G_{\bar{x}} \cap H_{\bar{x}}$.
- (ii) $\tilde{G}_{\bar{x}} \cap H_{\bar{x}} \neq \emptyset$ if and only if $\text{cl}(\tilde{G}_{\bar{x}} \cap H_{\bar{x}}) = G_{\bar{x}} \cap H_{\bar{x}}$.
- (iii) If, additionally, we suppose that the functions g_i , $i \notin I(\bar{x})$, are continuous, the functions h_j , $j = 1, \dots, m$, are continuously differentiable in a neighborhood of \bar{x} , and $\nabla h_j(\bar{x})$, $j = 1, \dots, m$, are linearly independent, then $\text{cl}(\tilde{G}_{\bar{x}} \cap H_{\bar{x}}) \subset T_{\bar{x}}$.

Assertions (i) and (ii) can be proved by reproducing the steps of the proof of Theorem 6.25 and Exercise 6.8. The proof of (iii) is also based on this theorem. More precisely, in a first stage, by using the hypothesis of linear independence of the gradients $\{\nabla h_j(\bar{x}), j = 1, \dots, m\}$, and applying the implicit function theorem, we have that the system $\{h_j(x) = 0, j = 1, \dots, m\}$ defines m variables as implicit functions of the others. Thus, the dimension of the space of the variables is reduced, while the new feasible set (in terms of the new variables) is expressed only by means of inequalities.

The content relations established in the previous theorem justify that the following conditions constitute *ad hoc* qualification hypotheses for our optimization problem with equality and inequality constraints (under appropriate assumptions of continuity and differentiability). It also ensures the existing chain of implications between them that we state below.

Constr. qualification	In brief	Hypothesis
Mangasarian–Fromovitz (also, Cottle)	MFCQ	$\{\nabla h_j(\bar{x}), j = 1, \dots, m\}$ linearly independent and $\tilde{G}_{\bar{x}} \cap H_{\bar{x}} \neq \emptyset$
Kuhn–Tucker	KTCQ	$T_{\bar{x}} = G_{\bar{x}} \cap H_{\bar{x}}$
Abadie	ACQ	$\mathcal{T}_{\bar{x}} = G_{\bar{x}} \cap H_{\bar{x}}$
Guignard	GCQ	$\text{cl}(\text{cone}\mathcal{T}_{\bar{x}}) = G_{\bar{x}} \cap H_{\bar{x}}$
Linear independence	LICQ	$\{\nabla g_i(\bar{x}), i \in I(\bar{x}); \nabla h_j(\bar{x}), j = 1, \dots, m\}$ linearly independent
Slater	SCQ	$\{\nabla h_j(\bar{x}), j = 1, \dots, m\}$ linearly independent, h_j are affine, $j = 1, \dots, m$, g_i are convex, $i \in I(\bar{x})$, $\exists x^0$ such that $h_j(x^0) = 0, j = 1, \dots, m$, and $g_i(x^0) < 0, i \in I(\bar{x})$

Suppose that the functions h_j , $j = 1, \dots, m$, are C^1 in a neighborhood of \bar{x} , the g_i , with $i \in I(\bar{x})$, are differentiable at \bar{x} , and the g_i , with $i \notin I(\bar{x})$, are continuous at \bar{x} , then:

$$\text{LICQ} \Rightarrow \text{MFCQ} \Rightarrow \text{KTCQ} \Rightarrow \text{ACQ} \Rightarrow \text{GCQ}. \quad \begin{matrix} \uparrow \\ \text{SCQ} \end{matrix} \quad (6.39)$$

The proof of the implication “ $\text{LICQ} \Rightarrow \text{MFCQ}$ ” comes from adapting to the new setting the arguments given in Proposition 6.27(i). More precisely, it follows from the following extension of Gordan Theorem 6.26, for systems including equalities: The system $\{a_i^T x < 0, i = 1, 2, \dots, s; a_i^T x = 0, i = s+1, \dots, r\}$ has no solution if and only if there are scalars $\lambda_1, \dots, \lambda_s \geq 0$, not all of them equal to zero, and μ_{s+1}, \dots, μ_r such that $\sum_{i=1}^s \lambda_i a_i + \sum_{i=s+1}^r \mu_i a_i = 0_n$.

Concerning the implication, “SCQ \Rightarrow MFCQ,” thanks to the convexity and differentiability of the functions g_i , $i \in I(\bar{x})$, we can write, for $i \in I(\bar{x})$,

$$0 > g_i(x^0) \geq g_i(\bar{x}) + \nabla g_i(\bar{x})^T(x^0 - \bar{x}) = \nabla g_i(\bar{x})^T(x^0 - \bar{x}),$$

and by the linearity of the functions h_j , we have, for $j = 1, 2, \dots, m$,

$$h_j(x^0) = h_j(\bar{x}) \quad \text{and} \quad \nabla h_j(\bar{x})^T(x^0 - \bar{x}) = 0.$$

Therefore, $d := x^0 - \bar{x} \in \tilde{G}_{\bar{x}} \cap H_{\bar{x}}$, and we are done.

6.4.1 Second-Order Optimality Conditions*

In this subsection, we consider the Lagrangian function associated with the problem (6.39), i.e., the function $L : \mathbb{R}^n \times \mathbb{R}_+^p \times \mathbb{R}^m \rightarrow \mathbb{R}$ given by

$$L(x, \lambda, \mu) := f(x) + \lambda^T g(x) + \mu^T h(x),$$

where g (respectively, h) represents the vector function having the g_i functions (respectively, the h_j) as coordinates. Moreover, we denote by $\nabla_x L(x, \lambda, \mu)$ the gradient of L with respect to x , i.e.,

$$\nabla_x L(x, \lambda, \mu) = \nabla f(x) + \sum_{i=1}^p \lambda_i \nabla g_i(x) + \sum_{j=1}^m \mu_j \nabla h_j(x),$$

and represent by $\nabla_{xx}^2 L(x, \lambda, \mu)$ the Hessian matrix of L with respect to x , that is

$$\nabla_{xx}^2 L(x, \lambda, \mu) := \nabla^2 f(x) + \sum_{i=1}^p \lambda_i \nabla^2 g_i(x) + \sum_{j=1}^m \mu_j \nabla^2 h_j(x),$$

where $\nabla^2 f(x)$, $\nabla^2 g_i(x)$, $i = 1, \dots, p$, $\nabla^2 h_j(x)$, $j = 1, \dots, m$, are the corresponding Hessian matrices.

In this way, the KKT conditions for problem (6.39) can alternatively be expressed as

$$\begin{aligned} \nabla_x L(\bar{x}, \bar{\lambda}, \bar{\mu}) &= 0_n, \\ \bar{\lambda}^T g(\bar{x}) &= 0, \quad \bar{\lambda} \geq 0_p, \\ g(\bar{x}) &\leq 0_p, \quad h(\bar{x}) = 0_m. \end{aligned} \tag{6.40}$$

In the statements in this subsection, we distinguish between two kinds of active constraints associated with the KKT point \bar{x} and the vector $\bar{\lambda}$ of KKT multipliers associated with the inequality constraints. We call *strongly active constraints* (also,

nondegenerate) to the constraints associated with the indices

$$I^+(\bar{x}, \bar{\lambda}) := \{i \in I(\bar{x}) : \bar{\lambda}_i > 0\},$$

while the remainder of the active inequality constraints are called *weakly active constraints*. In informal terms, this distinction is motivated by the fact that the latter class of active constraints plays no role in the KKT conditions (these conditions are verified even after eliminating such constraints).

Now, we introduce the closed convex cone associated with the pair $(\bar{x}, \bar{\lambda})$

$$M(\bar{x}, \bar{\lambda}) := \left\{ d \in \mathbb{R}^n \middle| \begin{array}{l} \nabla g_i(\bar{x})^T d \leq 0, i \in I(\bar{x}) \setminus I^+(\bar{x}, \bar{\lambda}), \\ \nabla g_i(\bar{x})^T d = 0, i \in I^+(\bar{x}, \bar{\lambda}), \\ \nabla h_j(\bar{x})^T d = 0, j = 1, 2, \dots, m. \end{array} \right\}.$$

We start by considering the possibility of having

$$M(\bar{x}, \bar{\lambda}) = \{0_n\},$$

which we call *Case 1*. In this case, every KKT point is a strict local minimum:

Theorem 6.37 (Second-order sufficient optimality condition, Case 1) *Suppose that $\bar{x} \in F$ is a KKT point for problem P in (6.39) with associated multipliers $\bar{\lambda} \in \mathbb{R}^p$, $\bar{\mu} \in \mathbb{R}^m$. If $M(\bar{x}, \bar{\lambda}) = \{0_n\}$, then \bar{x} is a strict local minimum of P .*

Proof Reasoning by contradiction, let us assume that \bar{x} is not a strict local minimum of P . Then, there exists a sequence $\{x_k\} \subset F \setminus \{\bar{x}\}$ converging to \bar{x} and such that

$$f(x_k) \leq f(\bar{x}), \quad k = 1, 2, \dots.$$

We can suppose without loss of generality that

$$d = \lim_{r \rightarrow \infty} \lambda_k(x_k - \bar{x}),$$

where $\lambda_k := \|x_k - \bar{x}\|^{-1}$. According to Theorem 6.36(i),

$$d \in T_{\bar{x}} \subset G_{\bar{x}} \cap H_{\bar{x}},$$

i.e., d satisfies

$$\nabla g_i(\bar{x})^T d \leq 0, \quad i \in I(\bar{x}), \quad \text{and} \quad \nabla h_j(\bar{x})^T d = 0, \quad j = 1, 2, \dots, m. \quad (6.41)$$

Actually, we can say something more. We shall prove that, in fact, $d \in M(\bar{x}, \bar{\lambda})$, which contradicts the assumption $M(\bar{x}, \bar{\lambda}) = \{0_n\}$, since $\|d\| = 1$.

If $\nabla g_{i_0}(\bar{x})^T d < 0$ for a certain $i_0 \in I^+(\bar{x}, \bar{\lambda})$, taking into account the KKT conditions

$$\nabla f(\bar{x}) + \sum_{i \in I^+(\bar{x}, \bar{\lambda})} \bar{\lambda}_i \nabla g_i(\bar{x}) + \sum_{j=1}^m \bar{\mu}_j \nabla h_j(\bar{x}) = 0_n,$$

and multiplying by d , we would get from (6.41) and $\nabla g_{i_0}(\bar{x})^T d < 0$ the inequality $\nabla f(\bar{x})^T d > 0$, but this inequality itself contradicts $f(x_k) \leq f(\bar{x})$, for all k . Certainly, we write

$$\nabla f(\bar{x})^T (x_k - \bar{x}) + o(\|x_k - \bar{x}\|) = f(x_k) - f(\bar{x}) \leq 0,$$

and

$$\nabla f(\bar{x})^T \frac{x_k - \bar{x}}{\|x_k - \bar{x}\|} + \frac{o(\|x_k - \bar{x}\|)}{\|x_k - \bar{x}\|} \leq 0, \quad k = 1, 2, \dots.$$

Taking limits, we reach the announced contradiction. \square

Next, we explore *Case 2*, i.e., the case where

$$\{0_n\} \subsetneq M(\bar{x}, \bar{\lambda}). \quad (6.42)$$

Theorem 6.38 (Second-order necessary optimality condition, Case 2) *Let \bar{x} be a local optimum of the problem P , introduced in (6.39). Suppose that $f, g_i, i \in I(\bar{x})$, and $h_j, j = 1, \dots, m$, are C^2 in a neighborhood of \bar{x} , and $g_i, i \notin I(\bar{x})$, are continuous functions at \bar{x} . Assume also that P satisfies any constraint qualification, entailing the existence of multipliers $\bar{\lambda} \in \mathbb{R}^p$ and $\bar{\mu} \in \mathbb{R}^m$ verifying (6.40) and (6.42). If, in addition, the related problem*

$$\begin{aligned} \hat{P} : \text{Min } & f(x) \\ \text{s.t. } & g_i(x) \leq 0, i \in I(\bar{x}) \setminus I^+(\bar{x}, \bar{\lambda}), \\ & g_i(x) = 0, i \in I^+(\bar{x}, \bar{\lambda}), \\ & h_j(x) = 0, j = 1, 2, \dots, m. \end{aligned}$$

satisfies the Abadie constraint qualification, then $\nabla_{xx}^2 L(\bar{x}, \bar{\lambda}, \bar{\mu})$ is positive semidefinite on $M(\bar{x}, \bar{\lambda})$.

Proof The proof is based on the observation that, if \bar{x} is a local minimum of P , it is also a minimum of the problem \hat{P} .

We distinguish with the symbol “ $\widehat{\cdot}$ ” the elements of problem \hat{P} . Thus, \widehat{F} is its feasible set, and $\widehat{T}_{\bar{x}}$, $\widehat{G}_{\bar{x}}$, and $\widehat{H}_{\bar{x}}$ represent, respectively, the cones of tangents at \bar{x} , the polar of the set of gradients at \bar{x} of the functions $g_i, i \in I(\bar{x}) \setminus I^+(\bar{x}, \bar{\lambda})$, and the orthogonal subspace to the gradients at \bar{x} of the functions involved in the equality constraints. With this notation, the cone $M(\bar{x}, \bar{\lambda})$ coincides with $\widehat{G}_{\bar{x}} \cap \widehat{H}_{\bar{x}}$. Moreover, since \hat{P} satisfies the Abadie constraint qualification, we have

$$M(\bar{x}, \bar{\lambda}) = \widehat{G}_{\bar{x}} \cap \widehat{H}_{\bar{x}} = \widehat{T}_{\bar{x}}. \quad (6.43)$$

Now, take $d \in M(\bar{x}, \bar{\lambda}) \setminus \{0_n\}$. From (6.43), $d = \lim_{k \rightarrow \infty} \rho_k(x_k - \bar{x})$, with ρ_k positive and $x_k \in \widehat{F}$ for all k , and $\{x_k\}$ converging to \bar{x} . The current differentiability assumptions allow us to write

$$\begin{aligned} f(x_k) &= f(\bar{x}) + \nabla f(\bar{x})^T(x_k - \bar{x}) \\ &\quad + \frac{1}{2}(x_k - \bar{x})^T \nabla^2 f(\bar{x})(x_k - \bar{x}) + o(\|x_k - \bar{x}\|^2), \\ g_i(x_k) &= g_i(\bar{x}) + \nabla g_i(\bar{x})^T(x_k - \bar{x}) \\ &\quad + \frac{1}{2}(x_k - \bar{x})^T \nabla^2 g_i(\bar{x})(x_k - \bar{x}) + o(\|x_k - \bar{x}\|^2), \quad i \in I^+(\bar{x}, \bar{\lambda}), \\ h_j(x_k) &= h_j(\bar{x}) + \nabla h_j(\bar{x})^T(x_k - \bar{x}) \\ &\quad + \frac{1}{2}(x_k - \bar{x})^T \nabla^2 h_j(\bar{x})(x_k - \bar{x}) + o(\|x_k - \bar{x}\|^2), \quad j = 1, \dots, m. \end{aligned} \tag{6.44}$$

Since

$$\nabla f(\bar{x}) + \sum_{i \in I^+(\bar{x}, \bar{\lambda})} \bar{\lambda}_i \nabla g_i(\bar{x}) + \sum_{j=1}^m \bar{\mu}_j \nabla h_j(\bar{x}) = 0_n,$$

from (6.44), we have

$$\begin{aligned} L(x_k, \bar{\lambda}, \bar{\mu}) &= f(x_k) + \sum_{i \in I^+(\bar{x}, \bar{\lambda})} \bar{\lambda}_i g_i(x_k) + \sum_{j=1}^m \bar{\mu}_j h_j(x_k) \\ &= f(\bar{x}) + \frac{1}{2}(x_k - \bar{x})^T \nabla_{xx}^2 L(x_k, \bar{\lambda}, \bar{\mu})(x_k - \bar{x}) + o(\|x_k - \bar{x}\|^2). \end{aligned} \tag{6.45}$$

Since $g_i(x_k) = h_j(x_k) = 0$ for all k , $i \in I^+(\bar{x}, \bar{\lambda})$ and $j = 1, 2, \dots, m$ (remember that $x_k \in \widehat{F}$, for all k), $f(x_k) \geq f(\bar{x})$ for k large enough (as \bar{x} is also a local minimum of \widehat{P}). Then, one has

$$0 \leq \frac{1}{2}(x_k - \bar{x})^T \nabla_{xx}^2 L(x_k, \bar{\lambda}, \bar{\mu})(x_k - \bar{x}) + o(\|x_k - \bar{x}\|^2),$$

for k large enough. Multiplying by ρ_k^2 and taking limits when $k \rightarrow +\infty$, we conclude that $d^T \nabla_{xx}^2 L(\bar{x}, \bar{\lambda}, \bar{\mu}) d \geq 0$. \square

Theorem 6.39 (Second-order sufficient optimality condition, Case 2) *Let $\bar{x} \in F$ be a KKT point for problem P in (6.39) with associated multipliers $\bar{\lambda} \geq 0_p$, $\bar{\mu} \in \mathbb{R}^m$ satisfying (6.42). Suppose that f , g_i , $i \in I(\bar{x})$, and h_j , $j = 1, \dots, m$, are \mathcal{C}^2 in a neighborhood of \bar{x} , and g_i , $i \notin I(\bar{x})$, are continuous at \bar{x} . If, additionally,*

$$\nabla_{xx}^2 L(\bar{x}, \bar{\lambda}, \bar{\mu}) \text{ is positive definite on } M(\bar{x}, \bar{\lambda}),$$

then \bar{x} is a strict local minimum of P .

Proof Reasoning by contradiction, assume that $d^T \nabla_{xx}^2 L(\bar{x}, \bar{\lambda}, \bar{\mu})d^T > 0$ for every $d \in M(\bar{x}, \bar{\lambda}) \setminus \{0_n\}$, but \bar{x} is not a strict local minimum of P . Then, there exists a sequence $\{x_k\} \subset F \setminus \{\bar{x}\}$ converging to \bar{x} and such that $f(x_k) \leq f(\bar{x})$, for all k ; thus,

$$L(x_k, \bar{\lambda}, \bar{\mu}) \leq f(\bar{x}), \quad \text{for all } k. \quad (6.46)$$

Moreover, we can assume without loss of generality that the sequence $\left\{ \frac{x_k - \bar{x}}{\|x_k - \bar{x}\|} \right\}$ converges to a certain vector $d \in \mathbb{R}^n \setminus \{0_n\}$. The reasoning in Theorem 6.37 allows us to assert that $d \in M(\bar{x}, \bar{\lambda})$.

Again by the differentiability assumptions, taking into account that \bar{x} is a KKT point, as well as (6.45) and (6.46), one gets

$$L(x_k, \bar{\lambda}, \bar{\mu}) = f(\bar{x}) + \frac{1}{2}(x_k - \bar{x})^T \nabla_{xx}^2 L(x_k, \bar{\lambda}, \bar{\mu})(x_k - \bar{x}) + o(\|x_k - \bar{x}\|^2) \leq f(\bar{x}).$$

This entails

$$\frac{1}{2}(x_k - \bar{x})^T \nabla_{xx}^2 L(x_k, \bar{\lambda}, \bar{\mu})(x_k - \bar{x}) + o(\|x_k - \bar{x}\|^2) \leq 0.$$

Dividing by $\|x_k - \bar{x}\|^2$ and letting $k \rightarrow +\infty$, we conclude that

$$d^T \nabla_{xx}^2 L(x_k, \bar{\lambda}, \bar{\mu})d \leq 0,$$

which contradicts the assumption, as $d \in M(\bar{x}, \bar{\lambda})$. Hence, \bar{x} is a strict local minimum of P . \square

The next theorem, whose proof is also based on the implicit function theorem, shows that, under appropriate assumptions, $\bar{\lambda}_i$ can be interpreted as the *price* (in the units of the objective function) that we would be willing to pay per unit increase in the right-hand side of the i th constraint (for small increments), as that unit would produce an improvement (decrease) of the objective function of approximately $\bar{\lambda}_i$ units.

Theorem 6.40 (Sensitivity theorem) *Let \bar{x} be a KKT point of the problem P in (6.39), with associated multipliers $\bar{\lambda} \geq 0_p$, $\bar{\mu} \in \mathbb{R}^m$. Suppose that f , g_i , $i \in I(\bar{x})$, and h_j , $j = 1, \dots, m$, are C^2 in a neighborhood of \bar{x} , and g_i , $i \notin I(\bar{x})$ are continuous at \bar{x} . Suppose further that the following conditions are fulfilled:*

(i) $\{\nabla g_i(\bar{x}), i \in I(\bar{x}); \nabla h_j(\bar{x}), j = 1, 2, \dots, m\}$ are linearly independent, i.e., LICQ holds.

(ii) $I(\bar{x}) = I^+(\bar{x}, \bar{\lambda})$ (all active constraints are strongly active).

(iii) $\nabla_{xx}^2 L(\bar{x}, \bar{\lambda}, \bar{\mu})$ is positive definite on the subspace $M(\bar{x}, \bar{\lambda})$.

Then, there exist neighborhoods $V \subset \mathbb{R}^n$ of \bar{x} and $W \subset \mathbb{R}^{p+m}$ of 0_{p+m} such that, for all $(\beta, \theta)^T \in W$, the parameterized problem

$$\begin{aligned} P(\beta, \theta) : \text{Min } & f(x) \\ \text{s.t. } & g(x) \leq \beta, \\ & h(x) = \theta, \end{aligned}$$

has a unique local optimum $\bar{x}(\beta, \theta)$ which is also strict; in particular $\bar{x}(0_p, 0_m) = \bar{x}$. In addition, $\bar{x}(\cdot, \cdot)$ is C^1 in W and

$$\nabla_{(\beta, \theta)} f(x(\beta, \theta))|_{(\beta, \theta)=(0_p, 0_m)} = -(\bar{\lambda}, \bar{\mu}).$$

6.5 Sequential Quadratic Programming Methods*

Sequential quadratic programming (SQP) is a family of very efficient methods for large constrained optimization problems with significant nonlinearities. SQP methods generate iterates by solving quadratic subproblems. Here, we present a local algorithm that motivates the SQP approach and that allows us to introduce the step computations in a simple setting.

Let us consider first the equality-constrained problem

$$\begin{aligned} P : \text{Min } & f(x) \\ \text{s.t. } & h_i(x) = 0, \quad i = 1, \dots, m, \end{aligned} \tag{6.47}$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $h_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, 2, \dots, m$ (equivalently, $h = (h_1, \dots, h_m)^T : \mathbb{R}^n \rightarrow \mathbb{R}^m$), $m \leq n$, all of them being, at least, of class C^2 .

Although problems containing only equality constraints are not frequent in practice, starting with the application of the SQP methodology to them provides the insight to better understand the design of SQP methods for problems with general constraints. The essential idea of SQP is to approximate (6.47) at the current iterate x_k by a quadratic optimization problem and to use its minimizer to define the new iterate x_{k+1} . This quadratic problem must be designed so that it yields a good step for the underlying constrained optimization problem P . The simple derivation of SQP methods, which we present next, is based on the application of Newton's method (recall Section 5.4) to the KKT optimality conditions for (6.47).

In Theorem 6.9, we derived necessary optimality conditions involving the Lagrangian function

$$L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i h_i(x) = f(x) + \lambda^T h(x),$$

with $\lambda = (\lambda_1, \dots, \lambda_m)^T$. More precisely, if \bar{x} is a local minimum of problem P , and $\nabla h(\bar{x}) = [\nabla h_1(\bar{x}) \mid \dots \mid \nabla h_m(\bar{x})]$ has full rank, then there is a unique vector of scalars $\bar{\lambda}$ such that $\nabla f(\bar{x}) + \nabla h(\bar{x})\bar{\lambda} = 0_n$; i.e., $(\bar{x}, \bar{\lambda})$ must be a solution of the system with $n + m$ equations and $n + m$ unknowns

$$F(x, \lambda) := \begin{pmatrix} \nabla f(x) + \nabla h(x)\lambda \\ h(x) \end{pmatrix} = \begin{pmatrix} 0_n \\ 0_m \end{pmatrix}. \tag{6.48}$$

Even for simple problems, solving the latter system of nonlinear equations is not an easy task (see Example 6.11). One possible approach consists of solving the nonlinear equations (6.48) by applying Newton's method. Taking into account that

$$\nabla F(x, \lambda) = \begin{bmatrix} \nabla_{xx}^2 L(x, \lambda) & \nabla h(x) \\ \nabla h(x)^T & 0_{m \times m} \end{bmatrix}$$

is symmetric, the pure Newton's method (5.46) for solving $F(x, \lambda) = 0_{n+m}$ is carried out by starting from some (x_0, λ_0) and then iterate

$$\begin{pmatrix} x_{k+1} \\ \lambda_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ \lambda_k \end{pmatrix} - (\nabla F(x_k, \lambda_k))^{-1} F(x_k, \lambda_k), \quad (6.49)$$

or, equivalently, $x_{k+1} = x_k + p_k$ and $\lambda_{k+1} = \lambda_k + \rho_k$, where p_k and ρ_k are solutions of the system

$$\begin{bmatrix} \nabla_{xx}^2 L(x_k, \lambda_k) & \nabla h(x_k) \\ \nabla h(x_k)^T & 0_{m \times m} \end{bmatrix} \begin{pmatrix} p \\ \rho \end{pmatrix} = - \begin{pmatrix} \nabla f(x_k) + \nabla h(x_k) \lambda_k \\ h(x_k) \end{pmatrix}. \quad (6.50)$$

It turns out that the second-order sufficient optimality conditions given in Theorem 6.15 imply local quadratic convergence for this Newton process. To prove such an assertion, we need the following lemma.

Lemma 6.41 *Let A be a symmetric $n \times n$ matrix and B a full-rank $n \times m$ matrix, with $m \leq n$, such that A is positive definite on the null space of B^T , i.e.,*

$$x^T A x > 0, \quad \forall x \neq 0_n \text{ such that } B^T x = 0_n.$$

Then, the $(n+m) \times (n+m)$ matrix

$$K := \begin{bmatrix} A & B \\ B^T & 0_{m \times m} \end{bmatrix}$$

is nonsingular.

Proof Reasoning by contradiction, if K is singular, there must exist $p \in \mathbb{R}^n$ and $q \in \mathbb{R}^m$ such that

$$\begin{pmatrix} p \\ q \end{pmatrix} \neq 0_{n+m} \quad \text{and} \quad \begin{bmatrix} A & B \\ B^T & 0_{m \times m} \end{bmatrix} \begin{pmatrix} p \\ q \end{pmatrix} = 0_{n+m}. \quad (6.51)$$

From the equality in (6.51), we get

$$Ap + Bq = 0_n, \quad (6.52)$$

$$B^T p = 0_m, \quad (6.53)$$

and from (6.53), it turns out that p belongs to the null space of B^T . Thus, multiplying (6.52) by p^T ,

$$p^T(Ap + Bq) = p^TAp + (B^Tp)^Tq = p^TAp = 0,$$

and, according to the assumption, it must be $p = 0_n$. Replacing p by 0_n in (6.52) gives rise to $Bq = 0_n$, but the full-rank assumption on B yields $q = 0_m$. We have concluded that $\begin{pmatrix} p \\ q \end{pmatrix}$ is null, in contradiction with the first assertion in (6.51). \square

Theorem 6.42 (Convergence of SQP methods) *Let $(\bar{x}, \bar{\lambda})$ satisfy the conditions of Theorem 6.15; i.e.,*

$$\begin{aligned} \nabla_x L(\bar{x}, \bar{\lambda}) &= 0_n, \quad h(\bar{x}) = 0_m, \\ y^T \nabla_{xx}^2 L(\bar{x}, \bar{\lambda})y &> 0, \quad \forall y \neq 0_n \text{ such that } \nabla h(\bar{x})^T y = 0_m, \end{aligned} \quad (6.54)$$

and assume also that $\nabla h(\bar{x})$ is full-rank. Then, for any starting point (x_0, λ_0) sufficiently close to $(\bar{x}, \bar{\lambda})$, the Newton iteration (6.49) converges quadratically to $(\bar{x}, \bar{\lambda})$, where \bar{x} is a strict local minimum of problem P .

Proof The current assumptions and Lemma 6.41 guarantee

$$\det \begin{bmatrix} \nabla_{xx}^2 L(\bar{x}, \bar{\lambda}) & \nabla h(\bar{x}) \\ \nabla h(\bar{x})^T & 0_{m \times m} \end{bmatrix} \neq 0.$$

Then, Newton's method applied to solve $F(x, \lambda) = 0_{n+m}$ converges quadratically, according to Theorem 5.19. \square

Let us make the following interesting observation: The Newton system (6.50) results from a quadratic-linear approximation of problem P . Indeed, (6.50) can be rewritten as

$$\begin{aligned} \nabla_{xx}^2 L(x_k, \lambda_k)p + \nabla h(x_k)(\rho + \lambda_k) &= -\nabla f(x_k) \\ \nabla h(x_k)^T p &= -h(x_k), \end{aligned}$$

and setting $\lambda := \lambda_k + \rho$, we get

$$\begin{aligned} \nabla_{xx}^2 L(x_k, \lambda_k)p + \nabla f(x_k) + \nabla h(x_k)\lambda &= 0_n, \\ \nabla h(x_k)^T p + h(x_k) &= 0_m, \end{aligned} \quad (6.55)$$

which coincides with the KKT system of the following approximation of P at x_k :

$$\begin{aligned} P(x_k, \lambda_k) : \text{Min}_p \frac{1}{2} p^T \nabla_{xx}^2 L(x_k, \lambda_k) p + \nabla f(x_k)^T p, \\ \text{s.t. } \nabla h(x_k)^T p + h(x_k) = 0_m. \end{aligned} \quad (6.56)$$

Since the constraints of this problem are affine, the KKT constraint qualification is satisfied (see Exercise 6.4). Moreover, under assumption (6.54), $P(x_k, \lambda_k)$ is a convex optimization problem if x_k is close enough to \bar{x} . Hence, solving the system (6.55) is equivalent to finding an optimal solution of $P(x_k, \lambda_k)$, and the Newton iteration (6.49) can be seen as the following sequential quadratic programming method:

SQP iteration: Given (x_k, λ_k) , compute the solution p_k to the quadratic problem $P(x_k, \lambda_k)$ in (6.56), with the corresponding multiplier $\lambda_{k+1} = \lambda_k + \rho_k$, and set $x_{k+1} = x_k + p_k$.

Observe that, in the absence of constraints, problem $P(x_k, \lambda_k)$ reduces to the unconstrained optimization problem

$$\text{Min}_p \frac{1}{2} p^T \nabla^2 f(x_k) p + \nabla f(x_k)^T p,$$

and the Newton step becomes

$$p_k = -(\nabla^2 f(x_k))^{-1} \nabla f(x_k),$$

in other words, we obtain the Newton direction; see (5.37).

Example 6.43 Let us apply the SQP method to the problem in Example 6.11, for which

$$\nabla f(x) = \begin{pmatrix} 2x_1 - x_2 + 1 \\ 2x_2 - x_1 \end{pmatrix} \quad \text{and} \quad \nabla h(x) = \begin{pmatrix} -3(1-x_1)^2 \\ -1 \end{pmatrix}.$$

We have

$$L(x, \lambda) = x_1^2 + x_2^2 - x_1 x_2 + x_1 + \lambda(1 - x_1)^3 - \lambda x_2,$$

so

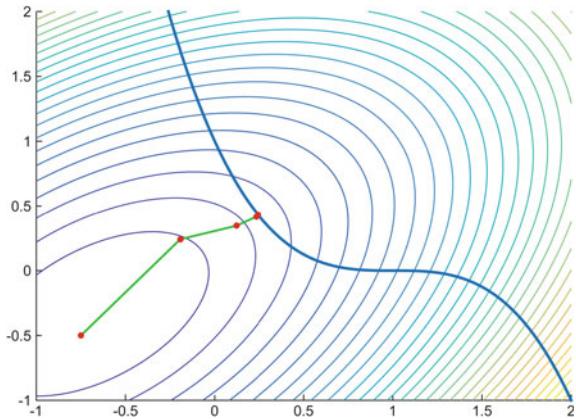
$$\nabla_{xx}^2 L(x, \lambda) = \begin{bmatrix} 2 + 6\lambda(1-x_1) & -1 \\ -1 & 2 \end{bmatrix}.$$

In Fig. 6.9, we show the SQP iteration for the starting point $((-3/4, -1/2)^T, 1)$. The sequence converges to $(\bar{x}, \bar{\lambda}) = ((0.2445258, 0.431183)^T, 0.6178347)$, which satisfies the first-order optimality condition (6.21).

Instead of performing a regular Newton step $x_{k+1} = x_k + p_k$, an alternative approach consists of using the solution p_k of $P(x_k, \lambda_k)$ as a search direction. In such a case, observe that if $h(x_k) = 0_m$ we see that $p = 0_m$ is feasible for $P(x_k, \lambda_k)$ and hence

$$\frac{1}{2} p_k^T \nabla_{xx}^2 L(x_k, \lambda_k) p_k + \nabla f(x_k)^T p_k \leq 0,$$

Fig. 6.9 The SQP iteration converges to the global minimum of the problem



entailing

$$\nabla f(x_k)^T p_k \leq -\frac{1}{2} p_k^T \nabla_{xx}^2 L(x_k, \lambda_k) p_k < 0,$$

provided that $\nabla_{xx}^2 L(x_k, \lambda_k)$ is positive definite, property which is held by continuity if assumption (6.54) holds and (x_k, λ_k) is close enough to $(\bar{x}, \bar{\lambda})$.

However, the SQP method does not necessarily produce iterates x_k that are feasible for P . Then, if $h(x_k) \neq 0_m$, then $p = 0_m$ is not feasible for $P(x_k, \lambda_k)$ and the optimal value of $P(x_k, \lambda_k)$ can be positive, in which case p_k need not to be a descent direction even when $\nabla_{xx}^2 L(x_k, \lambda_k)$ is positive definite (see Exercise 6.20).

Finally, we extend the SQP methodology to optimization problems in \mathbb{R}^n of the form

$$\begin{aligned} P : \text{Min } & f(x) \\ \text{s.t. } & h(x) = 0_m, \\ & g(x) \leq 0_p, \end{aligned}$$

with $m \leq n$, and this is achieved by generalizing, in a straightforward way, the approach based on the quadratic problems $P(x_k, \lambda_k)$. Given some (x_k, λ_k, μ_k) , we consider the quadratic minimization problem

$$\begin{aligned} P(x_k, \lambda_k, \mu_k) : \text{Min}_p & \frac{1}{2} p^T \nabla_{xx}^2 L(x_k, \lambda_k, \mu_k) p + \nabla f(x_k)^T p \\ \text{s.t. } & \nabla h(x_k)^T p + h(x_k) = 0_m, \\ & \nabla g(x_k)^T p + g(x_k) \leq 0_p. \end{aligned}$$

The general SQP method uses the solution p_k of $P(x_k, \lambda_k, \mu_k)$ as search direction for the next iteration $x_{k+1} = x_k + \alpha_k p_k$, with associated $\lambda_{k+1} \in \mathbb{R}^m$ and $\mu_{k+1} \geq 0_p$ satisfying the KKT system

$$\begin{aligned} \nabla_{xx}^2 L(x_k, \lambda_k, \mu_k) p + \nabla f(x_k) + \nabla h(x_k) \lambda + \nabla g(x_k) \mu &= 0_n, \\ \nabla h(x_k)^T p + h(x_k) &= 0_m, \quad \nabla g(x_k)^T p + g(x_k) \leq 0_p. \end{aligned}$$

6.6 Concluding Remarks*

Under certain conditions, the solutions to the sequence of *penalized problems* can be used to retrieve the KKT multipliers associated with the constraints of the original problem P in (6.4).

Assume that $\alpha(\cdot)$ is the penalty function introduced in (6.2) and (6.3). Additionally, suppose that ψ and ϕ are continuously differentiable and satisfy $\phi'(y) \geq 0$ for all y , and $\phi'(y) = 0$ for $y \leq 0$. Suppose, also, that the assumptions of Theorem 6.4 are fulfilled. Since x_η minimizes $f(x) + \eta\alpha(x)$, the gradient of this function must vanish at x_η , i.e.,

$$\nabla f(x_\eta) + \sum_{j=1}^m \eta \psi'(h_j(x_\eta)) \nabla h_j(x_\eta) + \sum_{i=1}^p \eta \phi'(g_i(x_\eta)) \nabla g_i(x_\eta) = 0_n. \quad (6.57)$$

Let \bar{x} be an accumulation point of the sequence $\{x_{\eta_k}\}$, $\eta_k \rightarrow \infty$. Without loss of generality, we can write

$$\lim_{k \rightarrow \infty} x_{\eta_k} = \bar{x}.$$

If $i \notin I(\bar{x}) = \{i : g_i(\bar{x}) = 0\}$, i.e., if $g_i(\bar{x}) < 0$, for k large enough it holds $g_i(x_{\eta_k}) < 0$, entailing $\eta_k \phi'(g_i(x_{\eta_k})) = 0$, due to the assumption made about ϕ' .

Now (6.57), with $\eta = \eta_k$ and k large enough, can be rewritten as follows:

$$0_n = \nabla f(x_{\eta_k}) + \sum_{j=1}^m v_k^j \nabla h_j(x_{\eta_k}) + \sum_{i \in I(\bar{x})} u_k^i \nabla g_i(x_{\eta_k}),$$

where v_k and u_k are those vectors whose components are

$$v_k^j := \eta_k \psi'(h_j(x_{\eta_k})), \quad j = 1, \dots, m, \quad (6.58)$$

$$u_k^i := \eta_k \phi'(g_i(x_{\eta_k})) \geq 0, \quad i \in I(\bar{x}). \quad (6.59)$$

If \bar{x} satisfies LICQ, there exist unique multipliers $\bar{\lambda}_j$, $j = 1, \dots, m$, and $\bar{\mu}_i \geq 0$, $i \in I(\bar{x})$, such that

$$0_n = \nabla f(\bar{x}) + \sum_{j=1}^m \bar{\lambda}_j \nabla h_j(\bar{x}) + \sum_{i \in I(\bar{x})} \bar{\mu}_i \nabla g_i(\bar{x}).$$

Since all the involved functions (f, h_j, g_i, ψ, ϕ) are continuously differentiable, for k large enough, the vectors

$$\{\nabla h_j(x_{\eta_k}), \quad j = 1, 2, \dots, m; \quad \nabla g_i(x_{\eta_k}), \quad i \in I(\bar{x})\}$$

are also linearly independent, and the associated scalars

$$\left\{ v_k^j, \ j = 1, \dots, m; \ u_k^i, \ i \in I(\bar{x}) \right\}$$

are unique as well. Then, it can be proved that

$$\begin{aligned} \bar{\lambda}_j &= \lim_{k \rightarrow \infty} \eta_k \psi'(h_j(x_{\eta_k})), \quad j = 1, \dots, m, \\ \bar{\mu}_i &= \lim_{k \rightarrow \infty} \eta_k \phi'(g_i(x_{\eta_k})), \quad i \in I(\bar{x}). \end{aligned}$$

Therefore, for sufficiently large k , the multipliers given in (6.58) and (6.59) can be used to estimate the KKT multipliers at the optimum point \bar{x} . For example, if α is the quadratic penalty function given by

$$\alpha(x) = \sum_{j=1}^m h_j^2(x) + \sum_{i=1}^p (g_i^+(x))^2,$$

that is, if

$$\begin{aligned} \psi(y) &= y^2 \Rightarrow \psi'(y) = 2y, \\ \phi(y) &= (y^+)^2 \Rightarrow \phi'(y) = 2y^+, \end{aligned}$$

then

$$\begin{aligned} \bar{\lambda}_j &= \lim_{k \rightarrow \infty} 2\eta_k h_j(x_{\eta_k}), \quad j = 1, \dots, m, \\ \bar{\mu}_i &= \lim_{k \rightarrow \infty} 2\eta_k g_i^+(x_{\eta_k}), \quad i \in I(\bar{x}). \end{aligned}$$

In particular, we realize that if $\bar{\mu}_i > 0$, for a certain $i \in I(\bar{x})$, then $g_i^+(x_{\eta_k}) > 0$ for k large enough, which means that the constraint $g_i(x) \leq 0$ is violated along the trajectory leading to \bar{x} .

Example 6.44 (Example 6.2 revisited) Recall that

$$x_{\eta_k} = \frac{\eta_k}{2\eta_k + 1} (1, 1)^T,$$

and we calculate

$$h(x_{\eta_k}) = -\frac{1}{2\eta_k + 1}.$$

Therefore,

$$v_k = 2\eta_k h(x_{\eta_k}) = -\frac{2\eta_k}{2\eta_k + 1},$$

and taking limits,

$$\bar{\lambda} = \lim_{k \rightarrow \infty} v_k = -1,$$

which is precisely the Lagrange multiplier associated with the optimal solution

$$\bar{x} = \lim_{k \rightarrow \infty} x_{\eta_k} = \frac{1}{2}(1, 1)^T.$$

Finally, let us consider the case in which f and $g_i, i \in I$, in problem (6.10) are differentiable functions and

$$\beta(x) := \gamma(g(x)),$$

where $g = (g_1, g_2, \dots, g_p)^T$ and $\gamma : \mathbb{R}^p \rightarrow \mathbb{R}$ is continuously differentiable at any point $y < 0_p$. Then,

$$\nabla \beta(x) = \sum_{i=1}^p \frac{\partial \gamma(g(x))}{\partial y_i} \nabla g_i(x),$$

and if x_k solves $P_B^{\eta_k}$, it holds

$$\nabla f(x_k) + \frac{1}{\eta_k} \sum_{i=1}^p \frac{\partial \gamma(g(x_k))}{\partial y_i} \nabla g_i(x_k) = 0_n. \quad (6.60)$$

Let us define

$$\lambda_i^k := \frac{1}{\eta_k} \sum_{i=1}^p \frac{\partial \gamma(g(x_k))}{\partial y_i}.$$

Then, (6.60) becomes

$$\nabla f(x_k) + \sum_{i=1}^p \lambda_i^k \nabla g_i(x_k) = 0_n.$$

Therefore, we can interpret the KKT vectors $\lambda^k := (\lambda_1^k, \lambda_2^k, \dots, \lambda_p^k)^T$ as a sort of vector of KKT multipliers. In fact, if f and $g_i, i \in I$, are also continuously differentiable functions and LICQ is satisfied at \bar{x} , it can be proved, using similar arguments to those used above, that $\lim_{k \rightarrow \infty} \lambda^k = \bar{\lambda}$, which is the (unique) KKT vector associated with \bar{x} .

6.7 Exercises

6.1 (a) Show that if $\bar{x} \in F$ is a local minimum on F of the differentiable function $f : \mathbb{R}^n \rightarrow \mathbb{R}$, where $F \subset \mathbb{R}^n$ is a convex set, then one must have:

$$\nabla f(\bar{x})^T(x - \bar{x}) \geq 0, \quad \forall x \in F. \quad (6.61)$$

(b) What consequence can be deduced from (6.61) if the set F is open?

(c) And what if F is an affine subspace?

(d) Show that condition (6.61) is also sufficient if f is also convex.

(e) Consider the quadratic and strictly convex function $f(x) = \frac{1}{2}x^T Q x - c^T x$. Let S be a vector subspace of \mathbb{R}^n , and suppose that the points $y, z \in \mathbb{R}^n$ are such that $z - y \notin S$. If

$$y^* = \operatorname{argmin}\{f(x) : x \in y + S\} \quad \text{and} \quad z^* = \operatorname{argmin}\{f(x) : x \in z + S\},$$

show, using part (c), that $z^* - y^*$ is Q -conjugate to any nonzero vector $v \in S$.

6.2 (a) Solve the problem

$$\begin{aligned} P : \operatorname{Min} & \sum_{i=1}^n x_i \\ \text{s.t.} & x_1 x_2 \dots x_n = 1, \\ & x_i > 0, \quad i = 1, 2, \dots, n. \end{aligned}$$

With the aim of avoiding the conditions $x_i > 0, i = 1, 2, \dots, n$, it is suggested to apply the change of variables

$$y_i = \ln(x_i), \quad i = 1, 2, \dots, n. \quad (6.62)$$

(b) Based on the result obtained, establish the well-known relationship between the arithmetic and geometric means for a set of positive numbers $x_i, i = 1, 2, \dots, n$,

$$(x_1 x_2 \dots x_n)^{1/n} \leq \frac{\sum_{i=1}^n x_i}{n}.$$

6.3 The well-known Cauchy–Schwarz inequality for the Euclidean norm asserts that

$$|x^T y| \leq \|x\| \|y\|, \quad \forall x, y \in \mathbb{R}^n. \quad (6.63)$$

This inequality is trivial if any of the two vectors is zero. Otherwise, inequality (6.63) is equivalent to

$$|x^T y| \leq 1, \quad \forall x, y \in \mathbb{R}^n \text{ with } \|x\| = \|y\| = 1. \quad (6.64)$$

Prove (6.64) by solving the optimization problem

$$\begin{aligned} P_y : \text{Max } & y^T x \\ \text{s.t. } & \|x\|^2 = 1, \end{aligned}$$

where y is an arbitrary vector of norm one.

6.4 Prove that KTCQ is satisfied at any feasible point of an optimization problem with affine constraints, that is, a problem with a system of constraints of the type

$$\{g_i(x) = a_i^T x - b_i \leq 0, i = 1, 2, \dots, m\}.$$

6.5 Consider the subset of \mathbb{R}^2

$$F := \{(t \cos(1/t), t \sin(1/t))^T, t > 0\} \cup \{(0, 0)\}.$$

Find the tangent cones $\mathcal{T}_{\bar{x}}$ and $T_{\bar{x}}$ to the set F at the point $\bar{x} = (0, 0)^T$.

6.6 Prove that if F is a convex set of \mathbb{R}^n and $\bar{x} \in F$, then

$$\mathcal{T}_{\bar{x}} = \text{clcone}(F - \bar{x}).$$

6.7 Characterize the tangent cone $\mathcal{T}_{\bar{x}}$ to F at $\bar{x} \in F = \{x \in \mathbb{R}^n : Ax = b\}$, where A is a real $m \times n$ matrix and $b \in \mathbb{R}^m$.

6.8 Show that $\text{cl}(\tilde{G}_{\bar{x}}) = G_{\bar{x}}$ if and only if $\tilde{G}_{\bar{x}} \neq \emptyset$.

6.9 Consider an optimization problem with constraints

$$g_i(x) \leq 0, i = 1, 2, \dots, m,$$

where the functions $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, 2, \dots, m$, are differentiable. Let F be the set of solutions of the latter system, and let $\bar{x} \in F$ be a point at which the Mangasarian–Fromovitz constraint qualification holds. Suppose that the functions $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i \in I(\bar{x})$, are also convex. Then, prove that the *Slater qualification* also holds at \bar{x} ; that is, there exists a point $\hat{x} \in \mathbb{R}^n$ such that

$$g_i(\hat{x}) < 0, \quad i \in I(\bar{x}).$$

6.10 Suppose that the functions $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, 2, \dots, p$, which appear in the constraints of an optimization problem

$$g_i(x) \leq 0, \quad i = 1, 2, \dots, p,$$

are all convex and differentiable, and that there exists a point \hat{x} such that

$$g_i(\hat{x}) < 0, \quad i = 1, 2, \dots, p.$$

For each point \bar{x} of the feasible set F , find the relationship existing between $\text{cone}(F - \bar{x})$ and the cone

$$G_{\bar{x}} = \{d \in \mathbb{R}^n : \nabla g_i(\bar{x})^T d \leq 0, i \in I(\bar{x})\}.$$

6.11 Let us consider the optimization problem

$$\begin{aligned} P : \text{Min } & f(x) \\ \text{s.t. } & Ax = b, \\ & g_i(x) \leq 0, i = 1, \dots, p, \end{aligned}$$

where $f : \mathbb{R}^n \rightarrow \mathbb{R}$, A is a real $m \times n$ matrix, $b \in \mathbb{R}^m$, and $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$, $i = 1, 2, \dots, p$. We also suppose that \bar{x} is a local minimizer for problem P , f and g_i are differentiable at \bar{x} , $i \in I(\bar{x})$, g_i are continuous at \bar{x} , $i \in \{1, 2, \dots, p\} \setminus I(\bar{x})$, and LICQ holds at \bar{x} , i.e.,

$\{\nabla g_i(\bar{x}), i \in I(\bar{x}); a_j^T, j = 1, 2, \dots, m\}$ is linearly independent,

where a_j denotes the j th row of A .

(a) Prove that the system

$$\begin{cases} \nabla f(\bar{x})^T d < 0, \\ \nabla g_i(\bar{x})^T d < 0, i \in I(\bar{x}), \\ Ad = 0_m, \end{cases} \quad (6.65)$$

has no solution d .

(b) Give a direct proof, based on (a) and the extended Gordan theorem, of the existence of vectors $\bar{\lambda} \in \mathbb{R}_+^p$ and $\bar{\mu} \in \mathbb{R}^m$ such that $(\bar{x}, \bar{\lambda}, \bar{\mu})$ satisfies the KKT conditions.

6.12 Determine the tangent cone $\mathcal{T}_{\bar{x}}$ to F at $\bar{x} \in F = \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$.

6.13 For any given vector y , we consider the problem

$$\begin{aligned} \text{Max } & y^T x \\ \text{s.t. } & x^T Q x \leq 1, \end{aligned} \quad (6.66)$$

where Q is a symmetric and positive definite matrix. Prove that the optimal value is $\sqrt{y^T Q^{-1} y}$, and use this result to prove the inequality

$$(y^T x)^2 \leq (y^T Q^{-1} y)(x^T Q x), \quad \forall x, y \in \mathbb{R}^n. \quad (6.67)$$

6.14 Apply the theory of Karush, Kuhn, and Tucker to characterize the norm of an $n \times n$ nonzero real matrix A , induced by the Euclidean norm, which is defined as

$$\|A\| := \max\{\|Ax\| : \|x\| \leq 1\}, \quad (6.68)$$

where $\|\cdot\|$ is the Euclidean norm in \mathbb{R}^n . Analyze the particular case where A is symmetric.

6.15 Consider the optimization problem in \mathbb{R}^2 given by

$$\begin{aligned} P : \text{Min } f(x) &= x_2 \\ \text{s.t. } x_1^2 + x_2^2 - 1 &\leq 0, \\ (x_1 - 2)^2 + x_2^2 - 1 &\leq 0. \end{aligned}$$

Answer the following questions:

- (a) Is there a (global) optimum \bar{x} ?
- (b) If so, does any constraint qualification hold at \bar{x} ?
- (c) Is \bar{x} a KKT point?

6.16 Consider the optimization problem in \mathbb{R}^2

$$\begin{aligned} P : \text{Min } f(x) &= x_1^2 - x_2^2 + x_1 \\ \text{s.t. } x_1^2 + x_2^2 &\leq 1, \\ x_1 &\geq 0, x_2 \geq 0. \end{aligned}$$

Check that $\bar{x} = 0_2$ is a KKT point, and nevertheless, it is not a local optimum. One can conclude, as a consequence of this fact, that the KKT conditions are not sufficient for optimality.

6.17 Consider the problem in \mathbb{R}^3

$$\begin{aligned} P : \text{Min } f(x) &= x_1^2 - x_2^2 + x_3 \\ \text{s.t. } x_2 &\geq 0, x_3 \geq 0. \end{aligned}$$

- (a) Prove that $\bar{x} = 0_3$ is a KKT point which is not a local minimum.
- (b) Check that the second-order necessary optimality condition is not satisfied.

6.18 Consider the problem in \mathbb{R}^2

$$\begin{aligned} P : \text{Min } f(x) &= (x_1 - x_2)^2 - x_1 + x_2 \\ \text{s.t. } x_1 - x_2 &\leq 0. \end{aligned}$$

- (a) Check that $\bar{x} = (1/2, 1/2)^T$ is a KKT point.
- (b) Does the second-order sufficient optimality condition hold at \bar{x} ?
- (c) Is \bar{x} a local optimum?

6.19 Consider the optimization problem in \mathbb{R}^2

$$\begin{aligned} P : \text{Min } f(x) &= (x_1 - 1)(x_2 - 1) \\ \text{s.t. } x_1 + x_2 &\leq 1, \\ x_1 &\geq 0, x_2 \geq 0. \end{aligned}$$

- (a) Find the set of KKT points of the problem P .
(b) Find all the minima.

6.20 Starting from $(x_0, \lambda_0) = ((1/2, 1/2)^T, -2)$, perform one step of the SQP method for the problem in \mathbb{R}^2 ,

$$\begin{aligned} P : \text{Min } f(x) &= x_1^2 + x_2^2 \\ \text{s.t. } h(x) &= x_1 - 1 = 0. \end{aligned}$$

Verify that the pure Newton's step p_0 is not a descent direction.

Correction to: Nonlinear Optimization



Correction to:

F. J. Aragón et al., *Nonlinear Optimization*,
Springer Undergraduate Texts
in Mathematics and Technology,
<https://doi.org/10.1007/978-3-030-11184-7>

The original version of the book was inadvertently published with incorrect reference numbers in the frontmatter. The erratum book has been updated with the change.

The updated version of the book can be found at
<https://doi.org/10.1007/978-3-030-11184-7>

Solutions to Selected Exercises

Problems of Chapter 1

1.2 Choosing an appropriate Cartesian reference, the points are $A(0, 0)$ and $B(3, 5)$, whereas the mountain range is $[1, 2] \times \mathbb{R}$. We have to decide the points $C(1, x)$ and $D(2, y)$ such that the cost of the road $[A, C, D, B]$ is minimized. Denoting $\alpha = \sqrt{1.6}$, the problem is

$$P : \text{Min } f(x, y) = \sqrt{1 + x^2} + \alpha\sqrt{1 + (y - x)^2} + \sqrt{1 + (5 - y)^2}.$$

As this function is differentiable and coercive, the global minimum will be the best of its critical points, that is $(2, 3)$. The optimal solution is $[A, C, D, B]$, with $C(1, 2)$ and $D(3, 3)$.

1.4 Denoting by x_1 , x_2 , and x_3 the length, the width, and the height of the box (in meters), the optimal solution to the approximating continuous model is $(2a, 2a, a/2)^T$, where $a = \sqrt[3]{5/2}$. The optimal choice is to take 115 trips, using a box with a square base of side 2.4052 and 0.6013 m height. The total cost of the operation will be 2885.41 c.u.

1.6 Let us denote by T the temperature and by t the time.

(a) $T = 25 - 20e^{-\alpha t}$.

(b) Defining $u = \ln(\frac{25-T}{20})$, the least squares fitting of the line $u = -\alpha t$ to the data provides $\alpha = -0.1552$, which should be replaced in (a).

1.7 The model is

$$\begin{aligned} P : \text{Min } & \sum_{i,j} w_{ij} \left\| \begin{pmatrix} x_i \\ y_i \end{pmatrix} - \begin{pmatrix} a_j \\ b_j \end{pmatrix} \right\|_1 \\ \text{s.t. } & \sum_{j=1}^n w_{ij} \leq c_i, i = 1, \dots, m, \\ & \sum_{i=1}^m w_{ij} \geq r_j, j = 1, \dots, n, \\ & w_{ij} \geq 0, \forall i \forall j. \end{aligned}$$

The objective function can be linearized by means of the technique introduced in Subsubsection 1.1.5.6 (“ ℓ_1 linear regression via linear optimization”).

1.10 After doing the variable transformation

$$\left. \begin{array}{l} x_1 = \sin y_1 \sin y_2 \dots \sin y_n \\ x_2 = \cos y_1 \sin y_2 \dots \sin y_n \\ x_3 = \cos y_2 \dots \sin y_n \\ \dots \\ x_n = \cos y_n \end{array} \right\},$$

one obtains an equivalent unconstrained problem with one less variable.

1.15 (a) By the first-order necessary condition, the unique candidate to local minima is, in all cases, $(1, -2)^T$.

(i) $(\bar{x}, \bar{y}) := (1, -2)$ satisfies the first-order sufficient condition because

$$\nabla f(x, y)^T \begin{pmatrix} x - 1 \\ y + 2 \end{pmatrix} = 2((x - 1) + (y + 2))^2 + 2(y + 2)^2 > 0 \quad (\text{A.1})$$

when $y \neq -2$ and also when $y = -2$, by the hypothesis that $(x, y) \neq (1, -2)$ (there is always a positive term in (A.1)). It cannot be asserted that $(1, -2)^T$ is a global minimum.

(ii) As $\nabla^2 f(x, y) = \begin{bmatrix} 2 & 2 \\ 2 & 6 \end{bmatrix}$ is positive definite, the unique singular point of f , $(1, -2)^T$, is a strict local minimum, but it cannot be guaranteed that it is a global minimum.

(iii) Since $(1, -2)^T$ is the unique candidate to be a local minimum, it will necessarily be a global minimum if one proves that f is coercive. Let us do the change of variables $x = u + 1$, $y = v - 2$. Then, we have

$$f(x, y) = g(u, v) = u^2 + 3v^2 + 2uv = (u + v)^2 + 2v^2 \geq 0, \quad \forall (u, v)^T \in \mathbb{R}^2.$$

Therefore, $S_\lambda(g) \neq \emptyset \Leftrightarrow \lambda \geq 0$. Obviously, $S_0(g) = \{0_2\}$. Let $\lambda > 0$. Then, we have $(u, v)^T \in S_\lambda(g) \Rightarrow |u + v| \leq \sqrt{\lambda}$ and $|v| \leq \sqrt{\frac{\lambda}{2}} \Rightarrow |u| \leq \sqrt{\lambda} + \sqrt{\frac{\lambda}{2}}$. Therefore, $S_\lambda(g)$ is bounded. As g is coercive on \mathbb{R}^2 , so is f .

(b) The point $(1, 1)^T$ is a strict local minimum, and it can only be classified by means of the first-order condition.

1.17 The key here is the coercivity of f , consequence of

$$S_\lambda(f) \subset \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3 : x, y, z \leq \ln \lambda, x + y + z \geq -\ln \left(\frac{\lambda}{2} \right) \right\},$$

for all $\lambda > 0$.

1.20 There is no global minimum, for any value of a , because $\lim_{x_1 \rightarrow -\infty} f(x_1, 0) = -\infty$; for $a \neq 0$, $(a, a)^T$ is a (strict) local minimum of f whereas there is no local minimum when $a = 0$.

1.22 Let x be the distance traveled by Bond, running along the beach shore, until he starts swimming. The time spent will be

$$f(x) = \frac{100 - x}{5} + \frac{1}{2}\sqrt{x^2 + 2500}$$

and we have to see if $\inf\{f(x) : x \in \mathbb{R}\} \leq 44$. As f is coercive and differentiable, the solutions will be critical points of f . The unique critical point is $\bar{x} = 21.82$ and as $f(\bar{x}) = 42.913 < 44$, Bond has enough time to disable the bomb and save the world.

1.23 If the position of a bather in $[a, b]$ is $Z \sim U(a, b)$ and the bar is on a , the average distance (round trip) of the bather to the bar is

$$E(2(Z - a)) = 2 \int_a^b \left(\frac{z - a}{b - a} \right) dz = b - a.$$

The expression is the same if the bar is on b . Therefore, the average distance of a bather who is served at an end point of the interval is its length. This observation will help us to express the objective functions.

(a) Let x be the distance from the bar to one of the ends of the beach, located, for example, at the point 0. The average sum of the distances of the bathers that are located on the left side of the bar is the average distance, x , multiplied by the number of bathers at that side of the beach, which is also x , as it is proportional to the length of the beach. Analogously, the total average distance walked around by the bathers that are located on the right side to the bar is $(1 - x)^2$. Therefore, we have to minimize the function $f(x) = x^2 + (1 - x)^2$ on $[0, 1]$. The unique optimal solution is $\frac{1}{2}$, with average distance walked around by the bathers of $\frac{1}{2}$ km, which is the same for both sides of the beach.

(b) It is sufficient to consider the bathers located in $[0, \frac{1}{2}]$, who will need to walk around an average distance of $\frac{1}{2}$ km.

(c) If x is the distance from the first bar to 0 (the nearest end of the beach) and y is the distance between both bars, the total distance is proportional to

$$f(x, y) = x^2 + \frac{y^2}{2} + (1 - x - y)^2,$$

whose minimum on the feasible set of the system $\{x + y \leq 1; x \geq 0; y \geq 0\}$ is $(\frac{1}{4}, \frac{1}{2})^T$, with $f(\frac{1}{4}, \frac{1}{2}) = \frac{1}{4}$ km (average distance). Bathers will need to walk around half of the distance than in (b).

Problems of Chapter 2

2.5 (a) f is convex and continuous, but it is not strongly convex, coercive, or Lipschitz continuous on \mathbb{R} (although it is on any bounded interval).

(b) f' is convex and Lipschitz continuous on \mathbb{R} (with constant 1), but it is not strongly convex or coercive.

2.6 (a) f_p is convex for $p = 1$ and when p is even. It is strongly convex for $p = 2$, and it is Lipschitz continuous at 0 for $p = 1$.

(b) f_p is convex for $p = 1$ and when p is even. It is strongly convex for $p = 2$, and it is Lipschitz continuous at 0, with Lipschitz constant $p\alpha^{p-1}$, for all $p \in \mathbb{N}$.

2.7 f_r satisfies the three properties for all $r \in \mathbb{N}$; the same happens with $\inf f_r = f_1$; finally, $\lim_r f_r = \sup_r f_r = |\cdot|$ is continuous and convex, but not differentiable.

2.9 C is convex, and $\nabla^2 f$ is positive semidefinite on $\text{int } C$. Moreover, f is continuous on C , and thus, it is convex on C .

2.12 (a) $C = \mathbb{R} \times \mathbb{R}_{--}$, where $\mathbb{R}_{--} = -\mathbb{R}_{++}$.

(b) As $f(0, y) = 0$ for all $y \in \mathbb{R}_{--}$, $\lim_{y \rightarrow -\infty} f(0, y) = 0$ and f is not coercive. Therefore, f is not strongly convex.

(c) If we take $y = -1$, we have

$$\lim_{x \rightarrow +\infty} \frac{\partial f(x, -1)}{\partial x} = \lim_{x \rightarrow +\infty} 2x \left(\frac{1}{e} + 1 \right) = +\infty,$$

so we suspect f is not Lipschitz continuous on horizontal directions. We shall use this idea for doing the proof: As

$$\begin{aligned} \lim_{r \rightarrow +\infty} \frac{f(r, -1) - f(1, -1)}{\|(r, -1) - (1, -1)\|} &= \lim_{r \rightarrow +\infty} \frac{r^2 \left(\frac{1}{e} + 1 \right) - \left(\frac{1}{e} + 1 \right)}{r - 1} \\ &= \left(\frac{1}{e} + 1 \right) \lim_{r \rightarrow +\infty} (r + 1) = +\infty, \end{aligned}$$

it does not exist $L \geq 0$ such that $|f(x, y) - f(u, v)| \leq L\|(x, y) - (u, v)\|$ for all $\begin{pmatrix} x \\ y \end{pmatrix}, \begin{pmatrix} u \\ v \end{pmatrix} \in C$. Therefore, f cannot be Lipschitz continuous on C .

2.14 (a) f is continuous, convex, and coercive, but, in general, it is neither strictly convex nor strongly convex (it is easy to prove it for $n = 1$).

(b) $F^* = [1, 3]$.

(c) 1.

2.17 $f \circ g$ is the composition of two convex functions and the second one is non-decreasing, so $f \circ g$ is convex.

2.19 (a) It is false: For instance, if $f(x, y) = x^2y^2$, one has that $\nabla^2 f(0, 0)$ is the null matrix, which is positive semidefinite, but $\nabla^2 f(x, y)$ is not positive semidefinite at points $\begin{pmatrix} x \\ y \end{pmatrix}$ that are arbitrarily close to the origin.

(b) It is true: If $\nabla^2 f(\bar{x})$ is positive definite, all the director principal minors are positive. Any principal minor is a sum of products of second-order partial derivatives of f , so it is a continuous function on \mathbb{R}^n . Therefore, each director principal minor is positive on an open ball with center \bar{x} . As there are n director principal minors, the intersection of the n balls is the ball with least radius, on which $\nabla^2 f(x)$ is positive definite and f is convex.

2.20 The argument would be correct if (a) was true, as well as the successive implications $(a) \Rightarrow (b), \dots, (d) \Rightarrow (e)$. This is not the case because $(d) \not\Rightarrow (e)$.

Problems of Chapter 3

3.5 (a) $\{x \in \mathbb{R}^2 : x_1 + x_2 = 0\}$.

$$(b) \bar{x} = \left(\frac{2617}{2439}, \frac{763}{813}, \frac{2792}{2439} \right)^T = (1.073, 0.9385, 1.1447)^T.$$

3.6 We have to minimize the function

$$h(x) := \|g - f\|_2^2 = \frac{1}{3}x_1^2 + x_1x_2 + x_2^2 - (\ln 2)x_1 - \frac{\pi}{2}x_2 + \int_0^1 \frac{dt}{(1+t^2)^2},$$

that is, $h(x) = \frac{1}{2}x^T Ax + a^T x + b$, with

$$A = \begin{pmatrix} \frac{2}{3} & 1 \\ 1 & 2 \end{pmatrix},$$

that is positive definite, $a = (-\ln 2, -\frac{\pi}{2})^T$ and $b = \int_0^1 \frac{dt}{(1+t^2)^2}$. Therefore, the unique optimal solution of P is the one of the linear system $\{Ax = -a\}$, that is, $\bar{x} = (6\ln 2 - \frac{3}{2}\pi, \pi - 3\ln 2)^T$. In conclusion, the affine function we are looking for is

$$g(t) = \left[6\ln 2 - \frac{3}{2}\pi \right]t + (\pi - 3\ln 2) \simeq 1.062 - 0.554t.$$

3.7 The point we are looking for is $\bar{x} = \frac{1}{m} \sum_{i=1}^m x^i$. When $m = 4$ and the points form a quadrilateral, \bar{x} is the midpoint of the midpoints of any pair of opposite edges. When the points form a parallelogram, there exist vectors $a, b, c \in \mathbb{R}^3$ such that $\{x^1, \dots, x^4\} = \{a, a+b, a+c, a+b+c\}$, with

$$\bar{x} = a + \frac{b+c}{2} = \frac{a+b}{2} + \frac{a+c}{2} = \frac{a}{2} + \frac{a+b+c}{2},$$

so \bar{x} is the intersection of the diagonals.

3.9 Denoting by S and p the area and the perimeter of the triangle, and by x, y, z the distances from P to the three sides, one has $S = \frac{1}{2}p(x + y + z)$, so that $x + y + z = \frac{2S}{p}$. Since $(xyz)^{\frac{1}{3}} \leq \frac{x+y+z}{3} = \frac{2S}{3p}$, the maximum is attained whenever $x = y = z$, which corresponds to the intersection of the three internal angle bisectors; that is, the incenter.

3.12 By the geometric-arithmetic inequality,

$$\begin{aligned} 1 &= 3\left(\frac{3x + 4y + 12z}{3}\right) \\ &\geq 3(3x)^{\frac{1}{3}}(4y)^{\frac{1}{3}}(12z)^{\frac{1}{3}} \\ &= 3\sqrt[3]{144}(xyz)^{\frac{1}{3}}, \end{aligned}$$

so the maximum is attained at $(\bar{x}, \bar{y}, \bar{z})^T = (\frac{1}{9}, \frac{1}{12}, \frac{1}{36})^T$ and the optimal value is $v(P) = \frac{1}{3^5 2^4} = \frac{1}{3888} = 2.572 \times 10^{-4}$.

3.13 As the volume and the height of both cans coincide, both cans have the same base area. So, it is sufficient to compare the lateral areas, S_c and S_r . We denote the volume and the height by V and h , while x, y are the dimensions of the rectangular base and z is the radius of the circular base. Taking into account that $V = \pi z^2 h$ and $S_c = 2\pi z h$ for the can with circular base, whereas $V = xyh$ and $S_r = 2(x + y)h$ for the one with rectangular base, and that $\sqrt{xy} \leq \frac{x+y}{2}$ by the geometric-arithmetic inequality, we have

$$\begin{aligned} S_r - S_c &= 2h(x + y - \pi z) \geq 2h\left(2\sqrt{xy} - \pi z\right) \\ &= 2h\left(2\sqrt{\frac{V}{h}} - \sqrt{\pi}\sqrt{\frac{V}{h}}\right) = 2\sqrt{hV}(2 - \sqrt{\pi}) > 0. \end{aligned}$$

3.15 We denote the base dimensions by x_1 and x_2 and the height by x_3 .

(a) The problem to solve is

$$\begin{aligned} P_1 : \text{Max } f_1(x) &= x_1 x_2 x_3 \\ \text{s.t. } h_1(x) &= x_1 x_2 + 2x_1 x_3 + 2x_2 x_3 = S \\ x &\in \mathbb{R}_{++}^n. \end{aligned}$$

By the geometric-arithmetic inequality, one has

$$\begin{aligned} S &= 3\left(\frac{x_1 x_2 + 2x_1 x_3 + 2x_2 x_3}{3}\right) \geq 3(x_1 x_2)^{\frac{1}{3}}(2x_1 x_3)^{\frac{1}{3}}(2x_2 x_3)^{\frac{1}{3}} \\ &= 3 \times 4^{\frac{1}{3}}(x_1 x_2 x_3)^{\frac{2}{3}} = 3 \times 4^{\frac{1}{3}}(f_1(x))^{\frac{2}{3}}. \end{aligned} \tag{A.2}$$

The maximum of $f_1(x)$ is attained when the inequality (A.2) is an equality, that is, when $x_1x_2 = 2x_1x_3 = 2x_2x_3 = \frac{S}{3}$. So, the optimal solution is $\bar{x} = \left(\sqrt{\frac{S}{3}}, \sqrt{\frac{S}{3}}, \frac{1}{2}\sqrt{\frac{S}{3}}\right)^T$.

(b) Now, the problem is

$$\begin{aligned} P_2 : \text{Min } f_2(x) &= x_1x_2 + 2x_1x_3 + 2x_2x_3 \\ \text{s.t. } h_2(x) &= x_1x_2x_3 = V \\ x &\in \mathbb{R}_{++}^n. \end{aligned}$$

By the geometric-arithmetic inequality, we have

$$f_2(x) = 3\left(\frac{x_1x_2 + 2x_1x_3 + 2x_2x_3}{3}\right) \geq 3 \times 4^{\frac{1}{3}}V^{\frac{2}{3}}.$$

The optimal solution $\bar{x} = \left(\sqrt[3]{2V}, \sqrt[3]{2V}, \frac{\sqrt[3]{2V}}{2}\right)^T$ is obtained by solving $x_1x_2 = 2x_1x_3 = 2x_2x_3 = 4^{\frac{1}{3}}V^{\frac{2}{3}}$.

3.16 We denote the base radio and the height by x_1 and x_2 , respectively.

(a) The problem to solve is

$$\begin{aligned} P_1 : \text{Max } f_1(x) &= x_1^2x_2 \\ \text{s.t. } h_1(x) &= c_1x_1^2 + c_2x_1x_2 = k_0, \\ x &\in \mathbb{R}_{++}^2, \end{aligned}$$

with $k_0 = \frac{c_0}{2\pi}$. By the geometric-arithmetic inequality, we have

$$\begin{aligned} h_1(x) &= 3\left(\frac{1}{3}c_1x_1^2 + \frac{1}{3}\frac{c_2x_1x_2}{2} + \frac{1}{3}\frac{c_2x_1x_2}{2}\right) \geq 3(c_1x_1^2)^{\frac{1}{3}}\left(\frac{c_2x_1x_2}{2}\right)^{\frac{2}{3}} \\ &= 3 \times 2^{-\frac{2}{3}}c_1^{\frac{1}{3}}c_2^{\frac{2}{3}}(f_1(x))^{\frac{2}{3}}, \end{aligned} \tag{A.3}$$

so the maximum value for $f_1(x)$ is obtained when $c_1x_1^2 = \frac{c_2x_1x_2}{2}$, that is, at

$$\bar{x} = \left(\sqrt{\frac{k_0}{3c_1}}, 2\sqrt{\frac{k_0c_1}{3c_2}}\right)^T.$$

(b) We have to solve the problem

$$\begin{aligned} P_2 : \text{Min } f_2(x) &= c_1x_1^2 + c_2x_1x_2 \\ \text{s.t. } h_2(x) &= x_1^2x_2 = \frac{V_0}{\pi} \\ x &\in \mathbb{R}_{++}^2. \end{aligned}$$

Now, inequality (A.3) can be rewritten as

$$f_2(x) \geq 3c_1^{\frac{1}{3}} \left(\frac{c_2}{2}\right)^{\frac{2}{3}} (h_2(x))^{\frac{2}{3}},$$

and the minimum cost is attained at

$$\bar{x} = \left(\sqrt[3]{\frac{c_2 V_0}{2c_1 \pi}}, \sqrt[3]{\frac{4c_1^2 V_0}{c_2 \pi}} \right)^T.$$

3.20 By applying the geometric-arithmetic inequality, we deduce that

$$f(x) = 5 \left(\frac{1}{5} \left(\frac{500}{x_1 x_2} \right) + \frac{1}{5} \left(\frac{500}{x_1 x_2} \right) + \frac{1}{5} (2x_1) + \frac{1}{5} (2x_2) + \frac{1}{5} (x_1 x_2) \right) \geq 5 \times 1000^{\frac{2}{5}},$$

where the equality is obtained if and only if

$$\frac{500}{x_1 x_2} = 2x_1 = 2x_2 = x_1 x_2. \quad (\text{A.4})$$

Since the system (A.4) is inconsistent, the lower bound given by the geometric-arithmetic inequality is never attained and we cannot solve the problem by means of the geometric-arithmetic inequality. Now, we consider the dual problem D in order to find the optimal solution of P . Since the dual feasible set G is formed by the solutions of the linear system

$$\begin{cases} y_1 + y_2 + y_3 + y_4 = 1 \\ -y_1 + y_2 + y_4 = 0 \\ -y_1 + y_3 + y_4 = 0 \end{cases},$$

with $y \in \mathbb{R}_{++}^4$, after some algebra, we obtain

$$G = \left\{ (y_1, 1 - 2y_1, 1 - 2y_1, 3y_1 - 1)^T : \frac{1}{3} < y_1 < \frac{1}{2} \right\}.$$

Therefore, D consists in maximizing

$$g(y) = \left(\frac{1}{3y_1 - 1} \right)^{3y_1 - 1} \left(\frac{2}{1 - 2y_1} \right)^{2(1 - 2y_1)} \left(\frac{1000}{y_1} \right)^{y_1}.$$

Taking $s = y_1$ to simplify and using the logarithmic transformation, D is equivalent to maximize

$$h(s) = (1 - 3s) \ln(3s - 1) + 2(1 - 2s)(\ln 2 - \ln(1 - 2s)) + s(\ln 1000 - \ln s)$$

on $\left] \frac{1}{3}, \frac{1}{2} \right[$. The maximum of h is attained at the unique point of the interval where the derivative vanishes, that is, approximately, the point $s = 0.438897$ corresponding to the point

$$\bar{y} = (0.4389, 0.1222, 0.1222, 0.3167)^T,$$

with $g(\bar{y}) = 84.82073$. After solving the system (3.22), we obtain the solution $\bar{x} = (5.18284, 5.18284)^T$.

3.23 The feasible set G of the dual problem is formed by the solutions on \mathbb{R}_{++}^4 of the system

$$\begin{cases} y_1 + y_2 + y_3 + y_4 = 1 \\ 3y_3 - y_4 = 0 \\ -6y_2 + 4y_4 = 0 \\ -y_1 + 2y_4 = 0 \\ 2y_2 + 2y_3 + 2y_4 = 0 \end{cases}.$$

As this system is inconsistent, $G = \emptyset$ and $F^* = \emptyset$.

Problems of Chapter 4

4.2 (a) The problem is

$$\begin{aligned} P_1 : \text{Min } f(x, y) &= x^3 + y^3 + (a - x - y)^3 \\ \text{s.t. } (x, y)^T &\in \mathbb{R}^2. \end{aligned}$$

As $f(r, r) \rightarrow -\infty$ when $r \rightarrow \infty$, $v(P_1) = -\infty$.

(b) Now the problem is

$$\begin{aligned} P_2 : \text{Min } f(x, y) &= x^3 + y^3 + (a - x - y)^3 \\ \text{s.t. } (x, y)^T &\in F_2, \end{aligned}$$

where

$$F_2 = \left\{ \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 : x > 0, y > 0, x + y < a \right\}$$

is an open convex set. As

$$\nabla^2 f(x, y) = 6 \begin{bmatrix} a - y & a - x - y \\ a - x - y & a - x \end{bmatrix}$$

is positive definite on F_2 , f is strictly convex on F_2 . Then, F_2^* is formed by the critical points of f on F , that is, $F_2^* = \left\{ \left(\frac{a}{3}, \frac{a}{3} \right)^T \right\}$. The optimal solution consists in decomposing a into three equal parts.

(c) In this case, the problem is a convex optimization problem with linear constraints:

$$\begin{aligned} P_3 : \text{Min } f(x, y) &= x^3 + y^3 + (a - x - y)^3 \\ \text{s.t. } x + y &\leq a \\ -x &\leq 0, -y \leq 0. \end{aligned}$$

Any $\binom{x}{y} \in F_3 = \text{cl } F_2$ can be written as $\binom{x}{y} = \lim_{r \rightarrow \infty} \binom{x_r}{y_r}$, with $\binom{x_r}{y_r} \in F_2$ for all $r \in \mathbb{N}$. Therefore, $f\left(\frac{a}{3}, \frac{a}{3}\right) \leq f(x_r, y_r)$ and taking limits when $r \rightarrow \infty$ we obtain $f\left(\frac{a}{3}, \frac{a}{3}\right) \leq f(x, y)$. The uniqueness is a consequence of that f is strictly convex.

4.4 Taking into account that the function $y \mapsto \frac{1}{y}$ is decreasing on \mathbb{R}_{++} , P is equivalent to the problem P_1 obtained by replacing the objective function f by $f_1(x_1, x_2) = -x_1 x_2$. Consider the relaxed problem P_2 obtained by replacing the strict inequalities $x_1 > 0$ and $x_2 > 0$ by the weak ones $-x_1 \leq 0$ and $-x_2 \leq 0$, whose unique minimal solution is $\bar{x} = (1, 1)^T$. Denoting by F_1 and F_2 the feasible sets of P_1 and P_2 , one has $\bar{x} \in F_1 \subset F_2$, so that \bar{x} is unique minimal solution of P_1 and, consequently, the unique minimal solution of P .

4.6 We want to know whether \bar{x} is the unique optimal solution of the problem

$$\begin{aligned} P : \text{Min } f(x) &= \|x\|^2 \\ \text{s.t. } x &\in F, \end{aligned}$$

where F represents the solution set of the system. P is a convex quadratic optimization problem with convex quadratic objective function f and linear inequality constraints, written in the form $g_i(x) \leq 0$, $i = 1, \dots, 6$ (with “ \leq ” instead of “ \geq ”). As f is strongly convex, P has a unique optimal solution. We check that $\bar{x} \in F$. Moreover, $I(\bar{x}) = \{2, 3\}$, with $\nabla f(\bar{x}) + \frac{3}{2}\nabla g_2(\bar{x}) + \frac{31}{6}\nabla g_3(\bar{x}) = 0_4$. Therefore, $F^* = \{\bar{x}\}$.

4.7 (a) It is obtained by doing the variable change $y_1 = 2x_1$ and $y_2 = x_2$; (b) $\bar{y} = (2, 2)^T$, that is, $\bar{x} = (1, 2)^T$; (d) $\bar{y} = (2, 3)^T$, that is, $\bar{x} = (1, 3)^T$.

4.11 The problem P is bounded because $f(x) = (x_1 - 1)^2 + x_2^2 - 1 \geq -1$. Moreover, P satisfies SCQ. As f is strongly convex (Proposition 3.1), $\bar{x} = \left(\frac{1}{2}, -\frac{1}{2}\right)^T$ is the unique optimal solution and $\bar{\lambda} = (1, 0)^T$ is the unique KKT vector. Therefore, the variation of the optimal value satisfies

$$\Delta \vartheta = \vartheta(z) - \vartheta(0_2) = \vartheta(z) - f\left(\frac{1}{2}, -\frac{1}{2}\right) = \vartheta(z) + \frac{1}{2} \geq -\bar{\lambda}^T z = -z_1,$$

for all $z \in \text{dom } \vartheta$. Thus, small perturbations of the right-hand side on the second constraint do not affect the optimal value (because it is not active at \bar{x}), whereas small perturbations of the right-hand side on the first constraint produce a minimum decreasing of the optimal value of one unit per each unit increased by the right-hand side.

4.15 We can consider any rectangular tetrahedron as the intersection of \mathbb{R}_+^3 with a half-space of the form $\frac{x}{a} + \frac{y}{b} + \frac{z}{c} \leq 1$. Its vertices are $O(0, 0, 0)$, $A(a, 0, 0)$,

$B(0, b, 0)$, and $C(0, 0, c)$. The hypotenuse area is a half of the norm of the cross product of $AB = (-a, b, 0)$ and $AC = (-a, 0, c)$, that is, $\frac{1}{2}\sqrt{a^2b^2 + a^2c^2 + b^2c^2}$. The pyramid height is the distance from the origin to the plane of the hypotenuse, $\frac{x}{a} + \frac{y}{b} + \frac{z}{c} - 1 = 0$, that is, $\frac{1}{\sqrt{\frac{1}{a^2} + \frac{1}{b^2} + \frac{1}{c^2}}} = h$, so we have that $\frac{1}{a^2} + \frac{1}{b^2} + \frac{1}{c^2} = \frac{1}{h^2} =: k$.

The problem we need to solve is

$$\begin{aligned} P_1 : \text{Min } f_1(a, b, c) &= a^2b^2 + a^2c^2 + b^2c^2 \\ \text{s.t. } &\frac{1}{a^2} + \frac{1}{b^2} + \frac{1}{c^2} = k, \\ &(a, b, c)^T \in \mathbb{R}_{++}^3. \end{aligned}$$

Taking $x_1 := k^{-1}a^{-2}$, $x_2 = k^{-1}b^{-2}$, and $x_3 = k^{-1}c^{-2}$, we obtain the following geometric optimization problem which is equivalent to P_1 :

$$\begin{aligned} P_2 : \text{Min } f_2(x_1, x_2, x_3) &= \frac{1}{x_1x_2} + \frac{1}{x_1x_3} + \frac{1}{x_2x_3} \\ \text{s.t. } &x_1 + x_2 + x_3 = 1, \\ &x \in \mathbb{R}_{++}^3. \end{aligned}$$

As we cannot solve P_2 by means of the geometric-arithmetic inequality, we consider the relaxed problem P_3 which is the result of eliminating the constraint set $x \in \mathbb{R}_{++}^3$. Applying KKT conditions, we obtain that the unique possible local minimum of P_3 is $\bar{x} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})^T$. As $\bar{x} \in \mathbb{R}_{++}^3$, this point is also the unique possible local (global) minimum of P_2 . Consequently, the unique possible global minimum of P_1 is obtained by taking $\bar{a} = \bar{b} = \bar{c} = \sqrt{\frac{3}{k}} = h\sqrt{3}$. Finally, we have to prove the existence of global minimum of P_1 . We shall see that F_1 is closed (so the sets $S_\lambda(f_1)$ are also closed) and that $S_\lambda(f_1)$ is bounded for all $\lambda > 0$.

Let $\{(a_r, b_r, c_r)^T\}_{r \in \mathbb{N}}$ be a sequence in F_1 such that $(a_r, b_r, c_r) \rightarrow (a, b, c)$. As $\frac{1}{a_r^2} + \frac{1}{b_r^2} + \frac{1}{c_r^2} = k$, $\frac{1}{a_r^2} \leq k$ and $a_r \geq \frac{1}{\sqrt{k}}$, so we have that $a \geq \frac{1}{\sqrt{k}} > 0$. Similarly, $b > 0$ and $c > 0$. Moreover, $\frac{1}{a^2} + \frac{1}{b^2} + \frac{1}{c^2} = k$ by a simple argument of continuity, so $(a, b, c)^T \in F_1$.

Let $(a, b, c)^T \in F_1$ be such that $f_1(a, b, c) \leq \lambda$, with λ a positive sufficiently large number. Then, one has $a, b, c \geq \frac{1}{\sqrt{k}}$ and $ab, ac, bc \leq \lambda$. Therefore, $\frac{1}{\sqrt{k}} \leq a \leq \frac{\lambda}{b} \leq \lambda\sqrt{k}$, inequalities also valid for b and c . So, the set $S_\lambda(f_1)$ is contained in the cube $\left[\frac{1}{\sqrt{k}}, \lambda\sqrt{k}\right]^3$.

4.16 We prove that P does not satisfy SCQ because $\|x\| - x_1 = 0$ for all $x \in \mathcal{F}(0) = F$.

(a) One has

$$\mathcal{F}(z) = \left\{ \begin{array}{ll} \emptyset, & z < 0 \\ \mathbb{R}_+ \times \{0\}, & z = 0 \\ \left\{ x \in \mathbb{R}^2 : x_1 \geq \frac{x_2^2}{2z} - \frac{z}{2} \right\}, & z > 0 \end{array} \right\} \text{ and } \vartheta(z) = \begin{cases} +\infty, & z < 0 \\ 1, & z = 0 \\ 0, & z > 0 \end{cases}.$$

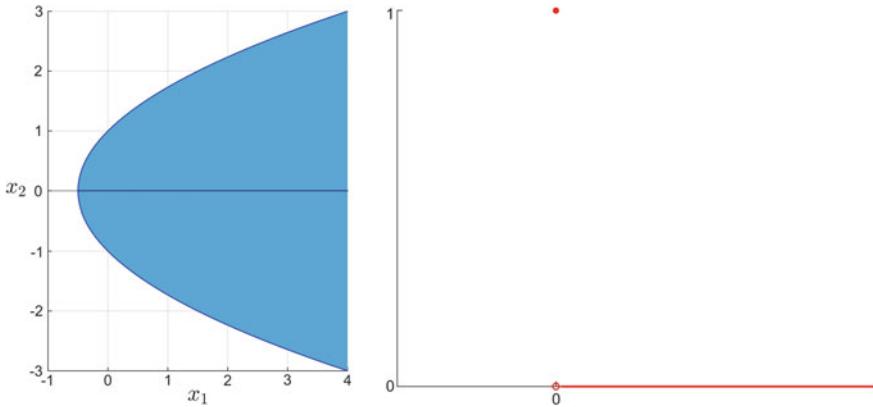


Fig. A.1 $\mathcal{F}(1)$ (left) y gph ϑ (right)

(b) ϑ is convex and differentiable on \mathbb{R}_{++} , but it is not continuous at 0, as it is shown in Fig. A.1. Also, observe that $0 \notin \text{int dom } \vartheta =]0, +\infty[$.

(c) There are no sensitivity vectors.

(d) As $L(x, 0) = e^{-x_2}$, $\inf_{x \in \mathbb{R}^2} L(x, 0) = 0 < 1 = v(P)$. If $\lambda > 0$,

$$\inf_{x \in \mathbb{R}^2} L(x, \lambda) \leq \inf_{z \in \mathbb{R}} L((z, 1)^T, \lambda) = \frac{1}{e} + \lambda \inf_{z \in \mathbb{R}} \left\{ \sqrt{z^2 + 1} - z \right\} = \frac{1}{e} < 1 = v(P).$$

Therefore, $v(D) \leq \frac{1}{e} < 1 = v(P)$.

4.17 It is easy to see that P satisfies SCQ.

(a) As the feasible set of $P(z)$ is $\mathcal{F}(z) = \{x \in \mathbb{R}^2 : x_1 + x_2 \leq z\}$, $\text{dom } \mathcal{F} = \mathbb{R}$,

$$\text{gph } \mathcal{F} = \left\{ \begin{pmatrix} z \\ x \end{pmatrix} \in \mathbb{R}^3 : x_1 + x_2 - z \leq 0 \right\} \text{ and } \vartheta(z) = \max \left\{ 0, -\frac{z}{\sqrt{2}} \right\},$$

see Fig. A.2.

(b) ϑ is continuous on \mathbb{R} and differentiable on $\mathbb{R} \setminus \{0\}$.

(c) The sensitivity vectors for P are the elements of the interval $[0, 1/\sqrt{2}]$.

(d) $F^* = \{0\}$.

(e) By Cauchy–Schwarz inequality, $|x_1 + x_2| \leq \sqrt{2}\|x\|$. Therefore, $x_1 + x_2 \geq -\sqrt{2}\|x\|$ and we have

$$L\left(x, \frac{1}{2}\right) = \|x\| + \frac{1}{2}(x_1 + x_2) \geq \|x\| - \frac{\sqrt{2}}{2}\|x\| = \left(1 - \frac{\sqrt{2}}{2}\right)\|x\| \rightarrow +\infty$$

when $\|x\| \rightarrow +\infty$. Thus, $L(x, \frac{1}{2})$ attains its global minimum on \mathbb{R}^2 . As $L(x, \frac{1}{2})$ has no critical points on $\mathbb{R}^2 \setminus \{0\}$ (where it is differentiable), its unique global minimum

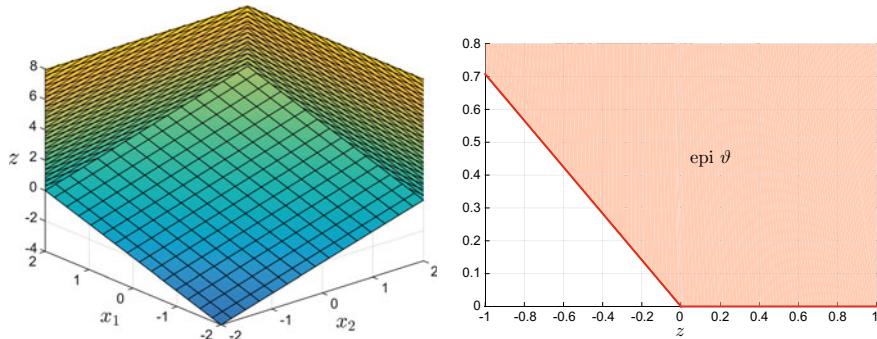


Fig. A.2 $\text{gph } \mathcal{F}$ (left) and $\text{gph } \vartheta$ (right)

is 0_2 and, effectively, $v(P) = \min_{x \in \mathbb{R}^2} L(x, \frac{1}{2}) = 0$ holds. Observe that $(0, 0, \frac{1}{2})$ is a saddle point.

(f) If $\lambda > \frac{1}{\sqrt{2}}$, then $L(x_1, x_1, \lambda) = \sqrt{2}(|x_1| + \sqrt{2}\lambda x_1) \rightarrow -\infty$ when $x_1 \rightarrow -\infty$. Therefore,

$$h(y) = \begin{cases} 0, & 0 \leq y \leq \frac{1}{\sqrt{2}}, \\ -\infty, & y > \frac{1}{\sqrt{2}}. \end{cases}$$

Then, we have $v(D) = v(P) = 0$, with $G^* = \left[0, \frac{1}{\sqrt{2}}\right]$.

4.18 (a) $F^* = \{(-1, -1, -1)^T\}$ (the metric projection of 0_3 onto the boundary of F ; (b) The value function is

$$\vartheta(z) = \begin{cases} \frac{(z-3)^2}{3}, & z < 3, \\ 0, & z \geq 3, \end{cases}$$

and the unique sensitivity vector (a scalar here) is $\bar{\lambda} = -\vartheta'(0) = 2$; (c) Here $Q = 2I_3$, $c = 0_3$, $A = (1, 1, 1)$, and $b = -3$.

Regarding the Lagrange dual of P , $L_Q(x, y) = \|x\|^2 + (x_1 + x_2 + x_3 + 3)y$, $h(y) = -\frac{3y^2}{4} + 3y$, and $v(D^L) = 3$, with unique optimal solution 2.

The Wolfe dual problem of P is

$$\begin{aligned} D_Q^W : \text{Max } & -\|u\|^2 + 3y \\ & u_i = -\frac{y}{2}, i = 1, 2, 3, \\ \text{s.t. } & y \geq 0, \end{aligned}$$

whose unique optimal solution is $(-1, -1, -1, 2)$, with $v(D_Q^W) = 3$.

Thus, strong duality holds for both dual pairs.

- 4.19** (a) The analytical solution cannot be obtained via KKT because P is convex but it does not satisfy SCQ. Nevertheless, $F^* = \{0\}$ because $F = \{0\}$.
 (b) The feasible set multifunction is

$$\mathcal{F}(z) = \begin{cases} \emptyset, & z < 0, \\ \{0\}, & z = 0, \\ [-\sqrt{z}, \sqrt{z}], & z > 0, \end{cases}$$

with

$$\text{gph } \mathcal{F} = \{(z, x) \in \mathbb{R}_+ \times \mathbb{R} : x^2 \leq z\} = \text{conv}\{(x^2, x) : x \in \mathbb{R}_+\}.$$

- (c) The value function is

$$\vartheta(z) = \begin{cases} +\infty, & z < 0, \\ -\sqrt{z}, & z \geq 0. \end{cases}$$

- (d) ϑ is differentiable on $\text{dom } \vartheta = \mathbb{R}_{++}$.
 (e) There does not exist any sensitivity vector (note that P is bounded, but it does not satisfy SCQ).
 (f) The Lagrange function is $L(x, \lambda) = x + \lambda x^2$, whose graph shown in Fig. A.3 suggests that there exist no saddle points. Indeed, let us suppose that $(\bar{x}, \bar{\lambda}) \in \mathbb{R} \times \mathbb{R}_+$ is a saddle point of L , that is,

$$\bar{x} + \lambda \bar{x}^2 \leq \bar{x} + \bar{\lambda} \bar{x}^2 \leq x + \bar{\lambda} x^2 \quad \forall x \in \mathbb{R}, \lambda \in \mathbb{R}_+.$$

From the first inequality, we deduce that $\bar{x} = 0$. Replacing it in the second one we have

$$0 \leq x + \bar{\lambda} x^2 \quad \forall x \in \mathbb{R},$$

which provides a contradiction, whether $\bar{\lambda} = 0$ or $\bar{\lambda} > 0$. Therefore, L has no saddle points.

- (g) There exist no KKT vectors.
 (h) Since SCQ fails, strong duality may fail too for both dual pairs.

On the one side, the Lagrange dual of P is

$$\begin{aligned} D^L : \text{Max } h(y) \\ \text{s.t. } y \geq 0, \end{aligned}$$

where

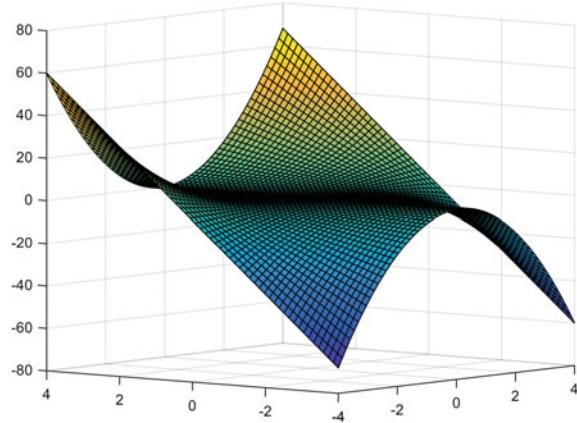
$$h(y) = \begin{cases} -\frac{1}{4y}, & y > 0, \\ -\infty, & \text{else,} \end{cases}$$

so that $v(D^L) = 0$, but the supremum of h is not attained on \mathbb{R}_+ .

On the other side, the Wolfe dual of P is

$$\begin{aligned} D^W : \text{Max } L(u, y) = u + yu^2 \\ \text{s.t. } y \geq 0, 1 + 2yu = 0, \end{aligned}$$

Fig. A.3 Graph of the Lagrange function is
 $L(x, \lambda) = x + \lambda x^2$



whose feasible set is the branch of hyperbola

$$G = \left\{ \left(u, -\frac{1}{2u} \right) \in \mathbb{R}^2 : u < 0 \right\}.$$

Since $L(u, -\frac{1}{2u}) = \frac{u}{2}$, $v(D^W) = 0$, but the supremum of L on G is not attained.

We conclude that the strong duality fails for both dual pairs, even though the duality gap is zero in both cases.

Problems of Chapter 5

5.2 (a) According with the computations in the proof of Proposition 5.17, one has that the stepsize when performing an exact line search at the point x_k in the direction $-g_k \equiv -\nabla f(x_k) = -Qx_k$ is

$$\alpha_k = \frac{g_k^T g_k}{g_k^T Q g_k}.$$

Thus, we obtain

$$g_0 = Q \left(\frac{1}{\lambda_{\min}} u_{\min} + \frac{1}{\lambda_{\max}} u_{\max} \right) = u_{\min} + u_{\max},$$

$$g_0^T g_0 = (u_{\min} + u_{\max})^T (u_{\min} + u_{\max}) = \|u_{\min}\|^2 + \|u_{\max}\|^2 = 2,$$

$$g_0^T Q g_0 = (u_{\min} + u_{\max})^T (\lambda_{\min} u_{\min} + \lambda_{\max} u_{\max}) = \lambda_{\min} + \lambda_{\max},$$

Therefore,

$$\alpha_0 = \frac{2}{\lambda_{\min} + \lambda_{\max}},$$

and

$$\begin{aligned}x_1 &= x_0 - \alpha_0 g_0 = \frac{1}{\lambda_{\min}} u_{\min} + \frac{1}{\lambda_{\max}} u_{\max} - \frac{2}{\lambda_{\min} + \lambda_{\max}} (u_{\min} + u_{\max}) \\&= \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \left(\frac{1}{\lambda_{\min}} u_{\min} - \frac{1}{\lambda_{\max}} u_{\max} \right).\end{aligned}$$

Hence, we deduce

$$\begin{aligned}f(x_1) &= \frac{1}{2} \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 \left(\frac{1}{\lambda_{\min}} u_{\min} - \frac{1}{\lambda_{\max}} u_{\max} \right)^T Q \left(\frac{1}{\lambda_{\min}} u_{\min} - \frac{1}{\lambda_{\max}} u_{\max} \right) \\&= \frac{1}{2} \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 \left(\frac{1}{\lambda_{\min}} u_{\min} - \frac{1}{\lambda_{\max}} u_{\max} \right)^T (u_{\min} - u_{\max}) \\&= \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 \frac{1}{2} \left(\frac{1}{\lambda_{\min}} + \frac{1}{\lambda_{\max}} \right) = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 f(x_0).\end{aligned}$$

Similar computations produce the following results

$$x_2 = x_1 - \alpha_1 g_1 = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 \left(\frac{1}{\lambda_{\min}} u_{\min} + \frac{1}{\lambda_{\max}} u_{\max} \right) = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 x_0,$$

and

$$f(x_2) = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 f(x_1) = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^4 f(x_0).$$

Completing the reasoning (by finite induction), it is possible to check that

$$x_k = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^k \left(\frac{1}{\lambda_{\min}} u_{\min} + \frac{(-1)^k}{\lambda_{\max}} u_{\max} \right), \quad k = 0, 1, 2, \dots,$$

deducing that

$$f(x_{k+1}) = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2 f(x_k) = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^{2k+2} f(x_0).$$

(b) Since

$$\frac{f(x_{k+1})}{f(x_k)} = \left(\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \right)^2,$$

the upper bound for the quotient $\frac{f(x_{k+1})}{f(x_k)}$ established in Proposition 5.17 is attained in this case.

(c) On one side the even points $x_0, x_2, \dots, x_{2p}, \dots$, and on the other side the odd points $x_1, x_3, \dots, x_{2p+1}, \dots$, are aligned with the origin, which is the unique global minimum of the function f .

5.5 (a) On one side, if $x \neq 0_n$, we have

$$\nabla f(x) = \frac{3}{2} \|x\|^{1/2} \nabla(\|x\|) = \frac{3}{2} \|x\|^{1/2} \frac{x}{\|x\|} = \frac{3}{2} \frac{x}{\|x\|^{1/2}},$$

and on the other side, computing the partial derivatives of f in $x = 0_n$, we obtain

$$\nabla f(0_n) = 0_n.$$

Therefore,

$$\|\nabla f(x)\| = \frac{3}{2} \|x\|^{1/2}, \quad \text{for all } x \in \mathbb{R}^n.$$

Any open set U containing the sublevel set $\{x \in \mathbb{R}^n : f(x) \leq K\}$ would contain certain ball $\rho\mathbb{B}$, since 0_n belongs to this sublevel set. If ∇f is Lipschitz continuous on U , there exists some constant $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in U. \quad (\text{A.5})$$

If we choose some x such that $\|x\| \leq \rho$ and $y = 0_n$, by (A.5), we get

$$\frac{3}{2} \|x\|^{1/2} \leq L\|x\|,$$

that is,

$$\frac{3}{2L} \leq \|x\|^{1/2},$$

but this inequality cannot hold if $\|x\|$ is sufficiently small.

(b) One has that

$$x_{k+1} = x_k - \alpha \nabla f(x_k) = \left(1 - \frac{3\alpha}{2} \frac{1}{\|x_k\|^{1/2}}\right) x_k, \quad k = 0, 1, \dots$$

Case 1.- Observe that if

$$1 - \frac{3\alpha}{2} \frac{1}{\|x_k\|^{1/2}} = 0 \Leftrightarrow \|x_k\| = \frac{9\alpha^2}{4},$$

it must be $x_{k+1} = 0_n$, and the method converges to the unique global minimum $\bar{x} = 0_n$ in a finite number of iterations.

Case 2.- If

$$1 - \frac{3\alpha}{2} \frac{1}{\|x_k\|^{1/2}} < -1 \Leftrightarrow \|x_k\| < \frac{9\alpha^2}{16},$$

then

$$\|x_{k+1}\| = \|x_k\| \left| 1 - \frac{3\alpha}{2} \frac{1}{\|x_k\|^{1/2}} \right| > \|x_k\|.$$

Case 3.- If

$$1 - \frac{3\alpha}{2} \frac{1}{\|x_k\|^{1/2}} > -1 \Leftrightarrow \|x_k\| > \frac{9\alpha^2}{16},$$

then

$$\|x_{k+1}\| < \|x_k\|.$$

Case 4.- Finally, if

$$1 - \frac{3\alpha}{2} \frac{1}{\|x_k\|^{1/2}} = -1 \Leftrightarrow \|x_k\| = \frac{9\alpha^2}{16},$$

then

$$x_{k+1} = -x_k,$$

and the algorithm infinitely oscillates between the two points x_k and $-x_k$, starting from the iteration k .

5.8 Since the matrix Q is invertible, all the eigenvalues $\lambda_1, \lambda_2, \dots, \lambda_n$ are different than zero, and we can consider the partition of the set of indices

$$N := \{i \mid \lambda_i < 0\} \quad \text{and} \quad P := \{i \mid \lambda_i > 0\}.$$

By hypothesis $N \neq \emptyset$.

Let $\{u_1, u_2, \dots, u_n\}$ be an orthonormal basis of associated eigenvectors. If x_0 is a starting point for the algorithm, we can express

$$x_0 = \sum_{i \in N} \xi_i u_i + \sum_{i \in P} \xi_i u_i.$$

We have

$$\begin{aligned} x_{k+1} &= x_k - \alpha Q x_k = (I - \alpha Q) x_k = \dots = (I - \alpha Q)^k x_0 \\ &= \sum_{i \in N} \xi_i (1 - \alpha \lambda_i)^k u_i + \sum_{i \in P} \xi_i (1 - \alpha \lambda_i)^k u_i. \end{aligned}$$

Taking norms

$$\|x_{k+1}\|^2 = \sum_{i \in N} (\xi_i)^2 (1 - \alpha \lambda_i)^{2k} + \sum_{i \in P} (\xi_i)^2 (1 - \alpha \lambda_i)^{2k}.$$

Case 1.- If there exists $i_0 \in N$ such that $\xi_{i_0} \neq 0$, we have

$$\lim_{k \rightarrow \infty} \|x_{k+1}\|^2 \geq \lim_{k \rightarrow \infty} (\xi_{i_0})^2 (1 - \alpha \lambda_{i_0})^{2k} = +\infty,$$

since $1 - \alpha \lambda_{i_0} > 1$, and the method diverges (neither the sequence $\{x_k\}$ nor any of its subsequences converge).

Case 2.- Otherwise, if we assume that $\xi_i = 0$ for all $i \in N$ and also $P \neq \emptyset$, and one takes α satisfying

$$\alpha < \min\{2/\lambda_i : i \in P\},$$

one has

$$|1 - \alpha \lambda_i| < 1, \quad \text{for all } i \in P,$$

and consequently

$$\lim_{k \rightarrow \infty} \|x_{k+1}\|^2 = \lim_{k \rightarrow \infty} \sum_{i \in P} \xi_i^2 (1 - \alpha \lambda_i)^{2k} = 0.$$

Thus, the sequence $\{x_k\}$ converges to 0_n , which is a saddle point of the function f .

5.9 Obviously, $\bar{x} = 0$ is the (strict global) minimum of f over \mathbb{R} .

(a) We have that

$$f'(x) = 2x^2 \operatorname{sign}(x) + x.$$

The function f is convex, since $f''(x) = 4|x| + 1 > 0$ for all $x \in \mathbb{R}$.

If we apply the steepest descent method with $\alpha_k = \frac{\gamma}{k+1}$, we obtain

$$x_{k+1} = x_k \left(1 - \frac{\gamma(2|x_k| + 1)}{k+1} \right), \quad \text{for } k = 0, 1, \dots \quad (\text{A.6})$$

Let $\gamma = 1$. If $x_k \geq k + 1$, from the expression above we deduce

$$x_{k+1} = x_k \left(1 - \frac{2x_k + 1}{k+1} \right) = \frac{x_k(k - 2x_k)}{k+1} \leq -(k+2).$$

If $x_k \leq -(k+1)$, we have

$$x_{k+1} = x_k \left(1 - \frac{-2x_k + 1}{k+1} \right) = \frac{x_k(k + 2x_k)}{k+1} \geq k+2.$$

We conclude that, for all k ,

$$|x_k| \geq k+1 \Rightarrow |x_{k+1}| \geq k+2.$$

Since $|x_0| \geq 1$, by induction, we obtain that $|x_k| \geq k+1$, for all k , and the sequence $\{x_k\}$ diverges.

(b) Define

$$y_k := |x_k|, \quad k = 1, 2, \dots.$$

Then, condition (5.101) becomes

$$\gamma(2y_0 + 1) < 2, \quad (\text{A.7})$$

and (A.6) implies

$$y_{k+1} = y_k \left| 1 - \frac{\gamma(2y_k + 1)}{k+1} \right|, \quad k = 0, 1, \dots. \quad (\text{A.8})$$

Based on (A.8), we will prove by induction that

$$\gamma(2y_k + 1) < 2, \quad k = 0, 1, \dots. \quad (\text{A.9})$$

Obviously (A.9) holds for $k = 0$ by the choice of x_0 (see (A.7)). Let us check that if (A.9) holds for some k , then it also holds for $k + 1$. In fact, (A.9) implies

$$1 - \frac{\gamma(2y_k + 1)}{k+1} > 1 - \frac{2}{k+1}. \quad (\text{A.10})$$

If $k = 0$, we deduce from (A.7) that

$$1 > 1 - \gamma(2y_0 + 1) > -1,$$

and by (A.8)

$$y_1 < y_0.$$

If $k > 0$, by (A.10) we get

$$1 > 1 - \frac{\gamma(2y_k + 1)}{k+1} \geq 0, \quad (\text{A.11})$$

and by (A.8),

$$y_{k+1} < y_k,$$

and also

$$\gamma(2y_{k+1} + 1) < \gamma(2y_k + 1) < 2.$$

Therefore, (A.9) holds for all k .

Finally, let us prove that $\{y_k\}$ converges to 0. By (A.8) and (A.11), we deduce that, for $k = 0, 1, \dots$,

$$y_{k+1} = y_k \left(1 - \frac{\gamma(2y_k + 1)}{k+1} \right) < y_k \left(1 - \frac{\gamma}{k+1} \right) < y_k \exp \left(-\frac{\gamma}{k+1} \right),$$

where we have used the inequality

$$1 - x < \exp(-x), \quad \text{for all } x > 0.$$

Thus,

$$\begin{aligned} y_{k+1} &< y_k \exp\left(-\frac{\gamma}{k+1}\right) < y_{k-1} \exp\left(-\frac{\gamma}{k+1} - \frac{\gamma}{k}\right) \\ &< y_0 \exp\left\{-\gamma\left(1 + \frac{1}{2} + \dots + \frac{1}{k+1}\right)\right\}, \end{aligned}$$

and then,

$$0 \leq \lim_{k \rightarrow \infty} y_{k+1} \leq y_0 \exp\left\{-\gamma \lim_{k \rightarrow \infty} \left(1 + \frac{1}{2} + \dots + \frac{1}{k+1}\right)\right\} = 0.$$

5.11 Having in mind that $\nabla f(x_k) = Qx_k$, we can write

$$x_{k+1} = (I - \alpha Q)x_k - \alpha e_k.$$

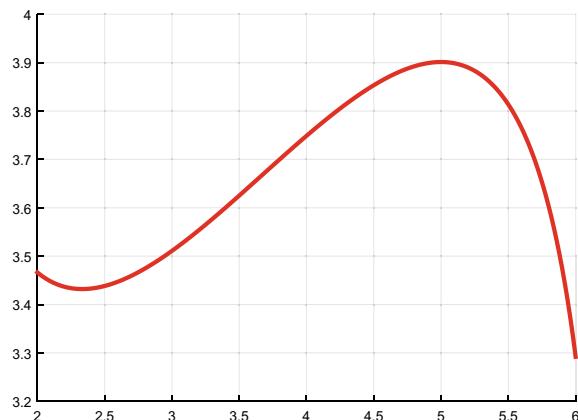
Hence,

$$\begin{aligned} \|x_{k+1}\| &\leq \|(I - \alpha Q)x_k\| + \|\alpha e_k\| \leq \max\{|1 - \alpha \lambda_{\min}|, |1 - \alpha \lambda_{\max}|\} \|x_k\| + \alpha \delta \\ &= q \|x_k\| + \alpha \delta \leq \dots \leq q^{k+1} \|x_0\| + \alpha \delta (1 + q + \dots + q^k). \end{aligned}$$

If $q < 1$, we obtain (5.102).

5.14 (a) We begin by plotting the graph of the function f over the interval $[2, 6]$, see Fig. A.4.

Fig. A.4 Graph of the function f



By making the derivative of f equal to zero,

$$f'(x) = 1 - \frac{1}{x-1} - \frac{1}{\frac{19}{3}-x} = 0$$

we obtain two solutions: $\bar{x} = \frac{7}{3}$, $\hat{x} = 5$. Since

$$f''(x) = \frac{1}{(x-1)^2} - \frac{1}{\left(\frac{19}{3}-x\right)^2},$$

one has that $f''(\bar{x}) = 0.5$, i.e., $\bar{x} = 7/3$ is a local minimum, and $f''(\hat{x}) = -0.5$, i.e., $\hat{x} = 5$ is a local maximum.

Further, $f''(\tilde{x}) = 0 \Rightarrow \tilde{x} = 11/3$. Therefore, the function f is convex on the interval $]1, 11/3[$.

(b) To apply the hint, we should check that $\varphi = f''$ is convex on $[2, 3] \subset]1, 11/3[$. Since

$$\varphi''(x) = f'''(x) = \frac{6}{(x-1)^4} - \frac{6}{(19/3-x)^4} \geq 0, \quad \forall x \in]1, 11/3[,$$

the convexity of φ is proved. Then, by (2.14), the Lipschitz constant of f'' on $[2, 3]$ is

$$L = \max \{|f''(2)|, |f''(3)|\}.$$

Since

$$f'''(x) = \frac{2}{(x-\frac{19}{3})^3} - \frac{2}{(x-1)^3},$$

we obtain $f'''(2) = -2.0246$ and $f'''(3) = -0.304$, and consequently, $L = 2.0246$.

(c) On one hand, it must be

$$\beta \leq \frac{7}{3} - 2 = \frac{1}{3}.$$

On the other hand, one has

$$\tilde{L} := L|1/f''(\bar{x})| = 2.0246 \times 2 = 4.0492,$$

and

$$\beta \tilde{L} < 1 \text{ if and only if } \beta < 1/(4.0492) = 0.24696.$$

Taking β as the minimum value between these two ($1/3$ and 0.24696), the convergence of the pure Newton's method with quadratic convergence rate is guaranteed if

$$x_0 \in [7/3 - 0.24696, 7/3 + 0.24696] = [2.0864, 2.5803].$$

(d) The pure Newton's method gives us the recursion formula

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} = x_k - \left(1 - \frac{1}{x_k - 1} - \frac{1}{\frac{19}{3} - x_k}\right) \frac{(x_k - 1)^2 \left(\frac{19}{3} - x_k\right)^2}{\left(\frac{22}{3} - 2x_k\right) \frac{16}{3}},$$

and if we take as initial point $x_0 = 2.5$, we obtain $x_1 = 2.3075$, $x_2 = 2.3328$, and $x_3 = 2.3333$.

5.16 (a) We have that

$$\begin{aligned} \varphi(\lambda_1, \dots, \lambda_n) &:= f\left(\sum_{i=1}^n \lambda_i d_i\right) = \frac{1}{2} \left(\sum_{i=1}^n \lambda_i d_i\right)^T Q \left(\sum_{i=1}^n \lambda_i d_i\right) - b^T \left(\sum_{i=1}^n \lambda_i d_i\right) \\ &= \sum_{i=1}^n \left(\frac{1}{2} (\lambda_i d_i)^T Q (\lambda_i d_i) - b^T (\lambda_i d_i) \right) \\ &= \sum_{i=1}^n \left(\frac{1}{2} \lambda_i^2 d_i^T Q d_i - \lambda_i b^T d_i \right). \end{aligned}$$

(b) Since this function is separable with respect to the λ_i , we can separately solve (i.e., using parallel computing) the problems

$$(P_i) \quad \min_{\lambda \in \mathbb{R}} \left\{ \frac{1}{2} (d_i^T Q d_i) \lambda^2 - (b^T d_i) \lambda \right\}.$$

(c) The objective function (of one variable λ) of the problem (P_i) is quadratic and strictly convex. Its minimum can be obtained by making the first derivative equal to zero:

$$(d_i^T Q d_i) \lambda_i^* = b^T d_i,$$

that is,

$$\lambda_i^* = \frac{b^T d_i}{d_i^T Q d_i}, \quad i = 1, 2, \dots, n,$$

and the optimal value is

$$v(P) = \sum_{i=1}^n \left(\frac{1}{2} (\lambda_i^*)^2 (d_i^T Q d_i) - \lambda_i^* (b^T d_i) \right) = -\frac{1}{2} \sum_{i=1}^n \frac{(b^T d_i)^2}{d_i^T Q d_i}.$$

5.19 The set $\{d_1, \dots, d_r\}$ is necessarily linearly dependent (otherwise, there would exist a basis of \mathbb{R}^n that spans \mathbb{R}^n positively). Hence, there are scalars $\lambda_1, \dots, \lambda_r$ (not all zero) such that $\lambda_1 d_1 + \dots + \lambda_r d_r = 0_n$. Take $i_0 \in \{1, \dots, r\}$ for which $\lambda_{i_0} \neq 0$; so

$$d_{i_0} = - \sum_{i=1, i \neq i_0}^r \frac{\lambda_i}{\lambda_{i_0}} d_i. \quad (\text{A.12})$$

Now let d be an arbitrary vector in \mathbb{R}^n . Since $\{d_1, \dots, d_r\}$ spans \mathbb{R}^n positively, there exist nonnegative scalars $\alpha_1, \dots, \alpha_r$ such that $d = \alpha_1 d_1 + \dots + \alpha_r d_r$. Then, using (A.12), one gets

$$d = \sum_{i=1, i \neq i_0}^r \alpha_i d_i + \alpha_{i_0} d_{i_0} = \sum_{i=1, i \neq i_0}^r \left(\alpha_i - \lambda_i \frac{\alpha_{i_0}}{\lambda_{i_0}} \right) d_i$$

Since d is arbitrary, we have proved that $\{d_1, \dots, d_r\} \setminus \{d_{i_0}\}$ spans \mathbb{R}^n .

Problems of Chapter 6

6.2 (a) Through the change of variables (6.62), we reformulate our optimization problem as follows:

$$\begin{aligned} \tilde{P} : \text{Min } & \sum_{i=1}^n \exp(y_i) \\ \text{s.t. } & \sum_{i=1}^n y_i = 0. \end{aligned}$$

The Lagrangian function of problem \tilde{P} is

$$L(y, \lambda) = \sum_{i=1}^n (\exp(y_i) + \lambda y_i).$$

The first-order optimality conditions are: There exist $\bar{y} \in \mathbb{R}^n$ and $\bar{\lambda} \in \mathbb{R}$ such that

$$\nabla_y L(\bar{y}, \bar{\lambda}) = 0_n \Leftrightarrow \exp(\bar{y}_i) + \bar{\lambda} = 0, \quad i = 1, 2, \dots, n, \quad (\text{A.13})$$

and

$$\nabla_\lambda L(\bar{y}, \bar{\lambda}) = 0 \Leftrightarrow \sum_{i=1}^n \bar{y}_i = 0. \quad (\text{A.14})$$

The system of $n + 1$ equations with $n + 1$ unknowns (A.13)–(A.14) has a *unique* solution

$$\bar{y}_i = 0, \quad i = 1, 2, \dots, n, \quad \bar{\lambda} = -1.$$

Since the objective function $\sum_{i=1}^n \exp(y_i)$ is continuous and coercive over the feasible set F of points y such that $\sum_{i=1}^n y_i = 0$ (if $y \in F$ and $\|y\| \rightarrow \infty$, one must have $\|y\|_\infty = \max_{i=1, 2, \dots, n} |y_i| = \max_{i=1, 2, \dots, n} y_i \rightarrow \infty$), the problem \tilde{P} must have a global minimum (which is unique, why?). Moreover, since every feasible point is regular (if $h(y) := \sum_{i=1}^n y_i$, $\nabla h(y) = (1, \dots, 1)^T$), the Lagrange necessary optimality conditions must be satisfied in such a minimum, which has to be $\bar{y} = 0_n$. Obviously, this minimum corresponds to the unique global minimum of P , which is

$$\bar{x}_i = \exp(\bar{y}_i) = 1, \quad i = 1, 2, \dots, n.$$

(b) Let x_1, \dots, x_n be some arbitrary positive numbers, and consider the product

$$a := x_1 x_2 \dots x_n.$$

Then

$$\left(\frac{x_1}{a^{1/n}} \right) \left(\frac{x_2}{a^{1/n}} \right) \dots \left(\frac{x_n}{a^{1/n}} \right) = 1,$$

and, by (a),

$$\sum_{i=1}^n \frac{x_i}{a^{1/n}} \geq n \text{ (optimal value of problem } P\text{).}$$

Consequently,

$$\frac{\sum_{i=1}^n x_i}{n} \geq a^{1/n} = (x_1 x_2 \dots x_n)^{1/n}.$$

6.3 Because the feasible set is compact, there exists a global maximum. Since the gradient of the function in the unique constraint is never zero on the feasible set, the Lagrange conditions must hold at \bar{x} . The Lagrange function is

$$L(x, \lambda) = y^T x + \lambda(\|x\|^2 - 1),$$

and the Lagrange conditions

$$y + 2\bar{\lambda}\bar{x} = 0_n, \| \bar{x} \| = 1.$$

Hence, since $2\bar{\lambda}\bar{x} = -y$, taking norms on both sides, we get

$$2|\bar{\lambda}| \|\bar{x}\| = 2|\bar{\lambda}| = \| -y \| = 1,$$

whence, $\bar{\lambda} = \pm 1/2$. Since

$$y^T \bar{x} = -2\bar{\lambda}(\bar{x})^T \bar{x} = -2\bar{\lambda},$$

if \bar{x} is a maximum, it must be $\bar{\lambda} = -1/2$, $y^T \bar{x} = 1$, and $\bar{x} = y$, leading to the well-known Cauchy–Schwarz inequality for the Euclidean norm.

6.5 We denote

$$x(t) := (t \cos(1/t), t \sin(1/t))^T, \quad \text{for } t > 0.$$

Obviously, $\|x(t)\| = t$, so every point of F can be identified by its norm, and $x(t) \rightarrow 0_2$ if and only if $t \rightarrow 0$.

(i) Let us prove first that $\mathcal{T}_{0_2} = \mathbb{R}^2$.

Since \mathcal{T}_{0_2} is a cone, it is sufficient to prove that if $d \in \mathbb{R}^2$ is a vector with norm one, then $d \in \mathcal{T}_{0_2}$. If d has norm one, there must exist some $\theta \in [0, 2\pi[$ such that $d \in (\cos \theta, \sin \theta)^T$. Consider now

$$t_r := \frac{1}{\theta + 2\pi r}, \quad r = 1, 2, \dots,$$

and define, for $r = 1, 2, \dots$,

$$\begin{aligned} x^r &:= x(t_r) := \frac{1}{\theta + 2\pi r}(\cos(\theta + 2\pi r), \sin(\theta + 2\pi r))^T \\ &= \frac{1}{\theta + 2\pi r}(\cos \theta, \sin \theta)^T = \frac{1}{\theta + 2\pi r}d^T. \end{aligned}$$

It is obvious that $F \ni x^r \rightarrow 0_2$ when $r \rightarrow \infty$, and that

$$d = \lim_{r \rightarrow \infty} \frac{1}{t_r}(x^r - 0_2),$$

that is, $d \in \mathcal{T}_{0_2}$.

(ii) Let us see now that $T_{0_2} = \{0_2\}$. In fact, if $d \in T_{0_2} \setminus \{0_2\}$, there must be some $\varepsilon > 0$ and some function $\alpha : [0, \varepsilon] \rightarrow F$ that is differentiable on $[0, \varepsilon]$ and such that $\alpha(0) = 0_2$ and $\alpha'(0) = d$.

Since $\alpha(\lambda) \in F$, the condition $\alpha'(0) = d$ implies that for $\lambda > 0$ sufficiently small, one has $\alpha(\lambda) \neq 0_2$; that is, there exists some $t_\lambda > 0$ such that $\alpha(\lambda) = x(t_\lambda)$. Since

$$\alpha'(0) = d = \lim_{\lambda \searrow 0} \frac{\alpha(\lambda)}{\lambda}, \tag{A.15}$$

and because differentiability implies continuity, one must have

$$\alpha(\lambda) = x(t_\lambda) \rightarrow \alpha(0) = 0_2 \quad \text{when } \lambda \rightarrow 0.$$

Thus, one must have $t_\lambda \rightarrow 0$ for $\lambda \rightarrow 0$, and by (A.15),

$$d = \lim_{\lambda \searrow 0} \frac{\alpha(\lambda)}{\lambda} = \lim_{\lambda \searrow 0} \frac{t_\lambda}{\lambda} \frac{x(t_\lambda)}{t_\lambda} = \lim_{\lambda \searrow 0} \frac{t_\lambda}{\lambda} (\cos(1/t_\lambda), \sin(1/t_\lambda))^T. \tag{A.16}$$

Taking norms, we obtain

$$\lim_{\lambda \searrow 0} \frac{t_\lambda}{\lambda} = \|d\| \neq 0,$$

and therefore, we deduce that (A.16) cannot hold because of the oscillating character of $(\cos(1/t_\lambda), \sin(1/t_\lambda))^T$ when $\lambda \searrow 0$.

6.6 Let $x \in F$ and let us define $d := x - \bar{x}$. If we consider the sequences

$$x^r := \bar{x} + (1/r)d \quad \text{and} \quad \lambda_r = r, \quad r = 1, 2, \dots,$$

by convexity of F , it is clear that $d \in \mathcal{T}_{\bar{x}}$. Since $x \in F$ has been arbitrarily chosen, one has $F - \bar{x} \subset \mathcal{T}_{\bar{x}}$. On the other hand, because $\mathcal{T}_{\bar{x}}$ is a cone, we have

$$\text{cone}(F - \bar{x}) = \mathbb{R}_+(F - \bar{x}) \subset \mathcal{T}_{\bar{x}},$$

and since $\mathcal{T}_{\bar{x}}$ is closed,

$$\text{cl cone}(F - \bar{x}) \subset \mathcal{T}_{\bar{x}}.$$

Let us prove the opposite inclusion. If $d \in \mathcal{T}_{\bar{x}}$, by definition of $\mathcal{T}_{\bar{x}}$, there exist sequences $\{x^r\} \subset F$ and $\{\lambda_r\} \subset]0, +\infty[$ such that

$$\lim_{r \rightarrow \infty} x^r = \bar{x} \quad \text{and} \quad \lim_{r \rightarrow \infty} \lambda_r(x^r - \bar{x}) = d.$$

Then, $x^r - \bar{x} \in F - \bar{x}$, and hence

$$\lambda_r(x^r - \bar{x}) \in \mathbb{R}_+(F - \bar{x}) = \text{cone}(F - \bar{x}),$$

which implies

$$d = \lim_{r \rightarrow \infty} \lambda_r(x^r - \bar{x}) \in \text{cl cone}(F - \bar{x});$$

that is, we have $d \in \text{cl cone}(F - \bar{x})$.

6.9 The Mangasarian–Fromovitz constraint qualification holds, that is, $\tilde{G}_{\bar{x}} \neq \emptyset$. Let $d \in \tilde{G}_{\bar{x}}$, i.e.,

$$\nabla g_i(\bar{x})^T d < 0, \quad i \in I(\bar{x}).$$

Then, for $i \in I(\bar{x})$, we have

$$g_i(\bar{x} + td) = g_i(\bar{x}) + t \nabla g_i(\bar{x})^T d + o(|t|) = t \left(\nabla g_i(\bar{x})^T d + \frac{o(|t|)}{t} \right),$$

and if one takes $t > 0$ sufficiently small

$$\nabla g_i(\bar{x})^T d + \frac{o(|t|)}{t} < 0,$$

which implies

$$g_i(\bar{x} + td) < 0. \tag{A.17}$$

Since $I(\bar{x})$ is finite, if $t_0 > 0$ is sufficiently small and we define $\hat{x} = \bar{x} + t_0 d$, by (A.17),

$$g_i(\bar{x}) < 0, \quad i \in I(\bar{x}),$$

that is, the Slater qualification holds.

6.10 The Slater qualification holds at every $\bar{x} \in F$, which implies the Abadie qualification, that is,

$$\mathcal{T}_{\bar{x}} = G_{\bar{x}}.$$

Since F is convex, we have by Exercise 6.6,

$$G_{\bar{x}} = \mathcal{T}_{\bar{x}} = \text{cl cone}(F - \bar{x}).$$

6.11 (a) Reasoning by contradiction, suppose that \bar{d} is a solution of (6.65). Then, for small t and thanks to the differentiability assumption, we can write

$$\begin{cases} f(\bar{x} + t\bar{d}) - f(\bar{x}) = t\left(\nabla f(\bar{x})^T \bar{d} + \frac{o(t)}{t}\right), \\ g_i(\bar{x} + t\bar{d}) - g_i(\bar{x}) = t\left(\nabla g_i(\bar{x})^T \bar{d} + \frac{o(t)}{t}\right), \quad i \in I(\bar{x}), \\ A(\bar{x} + t\bar{d}) - b = tA\bar{d} = 0_m, \end{cases}$$

and for t positive and small enough, taking into account the continuity of g_i , $i \in \{1, 2, \dots, p\} \setminus I(\bar{x})$ at \bar{x} , we see that $\bar{x} + t\bar{d}$ is feasible and $f(\bar{x} + t\bar{d}) < f(\bar{x})$, contradicting the local optimality of \bar{x} .

(b) Since the system (6.65) has no solution d , we apply the extended Gordan theorem (see Section 6.4) to conclude the existence of $\lambda_0 \geq 0$, $\hat{\lambda} \in \mathbb{R}_+^{I(\bar{x})}$, $\hat{\mu} \in \mathbb{R}^m$ such that $(\lambda_0, \hat{\lambda}) \neq 0_{1+|I(\bar{x})|}$ and

$$\lambda_0 \nabla f(\bar{x}) + \sum_{i \in I(\bar{x})} \hat{\lambda}_i \nabla g_i(\bar{x}) + A^T \hat{\mu} = 0_n. \quad (\text{A.18})$$

The possibility $\lambda_0 = 0$ is precluded because then $\hat{\lambda} \neq 0_{|I(\bar{x})|}$ and (A.18) would lead to a contradiction with LICQ. Hence, dividing (A.18) by $\lambda_0 > 0$ and defining

$$\bar{\lambda}_i := \begin{cases} \hat{\lambda}_i / \lambda_0, & \text{if } i \in I(\bar{x}), \\ 0, & \text{if } i \in \{1, 2, \dots, p\} \setminus I(\bar{x}), \end{cases} \quad \text{and} \quad \bar{\mu} := \frac{1}{\lambda_0} \hat{\mu},$$

we obtain

$$\nabla f(\bar{x}) + \sum_{i=1}^p \bar{\lambda}_i \nabla g_i(\bar{x}) + A^T \bar{\mu} = 0_n,$$

i.e. \bar{x} is a KKT point for problem P .

6.13 Consider the problem

$$\begin{aligned} P(y) : \text{Min } f(x) &= -y^T x \\ \text{s.t. } g(x) &= x^T Qx - 1 \leq 0, \end{aligned}$$

whose optimal value, that we represent by $v(y)$, is the opposite of the optimal value of problem (6.66), having both problems the same optimal solutions.

As the feasible set of $P(y)$ is compact and the objective function is linear, there is a global minimum. We will distinguish two cases:

(i) If $y = 0_n$, $v(0_n) = 0 = -\sqrt{0_n^T Q^{-1} 0_n}$ and (6.67) trivially holds for $y = 0_n$ and any x .

(ii) If $y \neq 0_n$, $\nabla f(x) = -y$ and $\nabla g(x) = 2Qx$.

Since Q is symmetric and positive definite, g is convex, $g(0_n) = -1$, and $\bar{x} = 0_n$ is a Slater point. Thus, if \bar{x} is a minimum of $P(y)$, there exists $\bar{\lambda} \geq 0$ satisfying the KKT conditions:

$$-y + 2\bar{\lambda}Q\bar{x} = 0_n \quad (\text{A.19})$$

and

$$\bar{\lambda}(\bar{x}^T Q \bar{x} - 1) = 0. \quad (\text{A.20})$$

As $y \neq 0_n$, obviously by (A.19) it cannot be $\bar{\lambda} = 0$, and the complementarity condition (A.20) implies

$$\bar{x}^T Q \bar{x} = 1. \quad (\text{A.21})$$

Then, we deduce from (A.19)

$$\bar{x} = \frac{1}{2\bar{\lambda}}Q^{-1}y. \quad (\text{A.22})$$

Premultiplying (A.19) by \bar{x}^T we obtain, having (A.21) into account,

$$-v(y) = y^T \bar{x} = 2\bar{\lambda} \bar{x}^T Q \bar{x} = 2\bar{\lambda},$$

and substituting into (A.22)

$$\bar{x} = \frac{1}{y^T \bar{x}} Q^{-1} y.$$

From this last relation, we have

$$y^T \bar{x} = \frac{1}{y^T \bar{x}} y^T Q^{-1} y,$$

and hence

$$(y^T \bar{x})^2 = y^T Q^{-1} y,$$

i.e.,

$$y^T \bar{x} = \pm \sqrt{y^T Q^{-1} y}.$$

Then, the optimal value of the original problem (maximization problem) is

$$-\nu(y) = y^T \bar{x} = \sqrt{y^T Q^{-1} y}.$$

If $x \neq 0_n$, it is obvious that

$$\tilde{x} := \frac{1}{\sqrt{x^T Q x}} x$$

is a feasible point, so

$$y^T \left(\frac{1}{\sqrt{x^T Q x}} x \right) \leq \sqrt{y^T Q^{-1} y},$$

which is the same as (6.67).

6.14 The optimization problem (6.68) is equivalent to the problem

$$\begin{aligned} P : \text{Min } f(x) &= -\frac{1}{2} \|Ax\|^2 = -\frac{1}{2} x^T A^T A x, \\ \text{s.t. } g(x) &= \frac{1}{2} x^T x - \frac{1}{2} \leq 0. \end{aligned}$$

This problem has an optimal solution because we are minimizing a continuous function over a compact set. Let us see that if \bar{x} is an optimal solution of P , one has $\|\bar{x}\| = 1$:

(a) If $\bar{x} = 0_n$, by (6.68), we have $\|A\| = 0$, and because the norm is zero, A must be the null matrix, which can be dismissed by the standing assumptions.

(b) If $\|A\| > 0$ and $0 < \|\bar{x}\| < 1$, it is clear that $\bar{z} := \|\bar{x}\|^{-1} \bar{x}$ satisfies

$$-\frac{1}{2} \bar{z}^T A^T A \bar{z} = -\frac{1}{2 \|\bar{x}\|^2} \bar{x}^T A^T A \bar{x} < -\frac{1}{2} \bar{x}^T A^T A \bar{x},$$

which contradicts that \bar{x} is an optimal solution of P .

Therefore, $\|\bar{x}\| = 1$ and $\nabla g(\bar{x}) = \bar{x} \neq 0_n$; that is, the linear independence qualification is satisfied at \bar{x} . Then, the KKT conditions hold, and there exists $\bar{\lambda} \geq 0$ such that

$$-A^T A \bar{x} + \bar{\lambda} \bar{x} = 0_n.$$

Obviously, one must have $\bar{\lambda} > 0$, because $\bar{\lambda} = 0$ implies $-A^T A \bar{x} = 0_n$, and then $f(\bar{x}) = -\frac{1}{2} \|A\|^2 = 0$, a possibility which is excluded.

From $A^T A \bar{x} = \bar{\lambda} \bar{x}$, we deduce that \bar{x} is an eigenvector associated with the eigenvalue $\bar{\lambda} > 0$. Now, let $\tilde{\lambda}$ be another eigenvalue (which is real, since $A^T A$ is symmetric) and let \tilde{x} be an associated eigenvector of norm one. By optimality of \bar{x} , one has

$$\tilde{\lambda} = \tilde{x}^T (\tilde{\lambda} \tilde{x}) = \tilde{x}^T A^T A \tilde{x} \leq \bar{x}^T A^T A \bar{x} = \bar{x}^T (\bar{\lambda} \bar{x}) = \bar{\lambda},$$

i.e., $\bar{\lambda}$ is the largest eigenvalue of $A^T A$.

Thus,

$$\|A\|^2 = \bar{x}^T A^T A \bar{x} = \bar{\lambda},$$

and

$$\|A\| = \sqrt{\rho(A^T A)},$$

whence, $\rho(A^T A)$ is the spectral radius of the matrix $A^T A$, that is, the maximum eigenvalue of the matrix $A^T A$.

If A is symmetric and $\mu_{min} := \mu_1 \leq \mu_2 \leq \dots \leq \mu_n := \mu_{max}$ are its eigenvalues, it must hold

$$\|A\| = \sqrt{\rho(A^2)} = \sqrt{\max\{\mu_{min}^2, \mu_{max}^2\}} = \max\{|\mu_{min}|, |\mu_{max}|\}.$$

6.19 (a) Simple computations show that there are three KKT points:

$$\begin{aligned}\bar{x} &= (1/2, 1/2)^T, \quad I(\bar{x}) = \{1\}, \quad \bar{\lambda} = 1/2, \\ \tilde{x} &= (0, 1)^T, \quad I(\tilde{x}) = \{1, 2\}, \quad \tilde{\lambda}_1 = \tilde{\lambda}_2 = 1, \\ \hat{x} &= (1, 0)^T, \quad I(\hat{x}) = \{1, 3\}, \quad \hat{\lambda}_1 = \hat{\lambda}_3 = 1.\end{aligned}$$

(b) As the constraints are linear, the KTCQ must hold (see Exercise 6.4), and then every local optimum of P will be a KKT point. Therefore, the local optima can be found among all the KKT points, but not every KKT point will be a local optimum.

We have

$$\begin{aligned}L(x, \lambda) &= (x_1 - 1)(x_2 - 1) + \lambda_1(x_1 + x_2 - 1) + \lambda_2(-x_1) + \lambda_3(-x_2), \\ \nabla_x L(x, \lambda) &= (x_2 - 1 + \lambda_1 - \lambda_2, x_1 - 1 + \lambda_1 - \lambda_3)^T,\end{aligned}$$

and

$$\nabla_{xx}^2 L(x, \lambda) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

(b1) For $\bar{x} = (1/2, 1/2)^T$, one has

$$M(\bar{x}, \bar{\lambda}) = \{d \in \mathbb{R}^2 : \nabla g_1(\bar{x})^T d = 0\} = \{(d_1, d_2)^T : d_1 + d_2 = 0\},$$

and for $d_0 = (-1, 1)^T \in M(\bar{x}, \bar{\lambda})$ it holds

$$d_0^T \nabla_{xx}^2 L(\bar{x}, \bar{\lambda}) d_0 = -2,$$

and the second-order necessary optimality condition fails; that is, $\bar{x} = (1/2, 1/2)^T$ is not a local minimum.

(b2) For $\tilde{x} = (0, 1)^T$, one has

$$\begin{aligned}M(\tilde{x}, \tilde{\lambda}) &= \{d \in \mathbb{R}^2 : \nabla g_1(\tilde{x})^T d = 0, \nabla g_2(\tilde{x})^T d = 0\} \\ &= \{(d_1, d_2)^T : d_1 + d_2 = 0 \text{ and } d_2 = 0\} = \{0_2\},\end{aligned}$$

and the second-order necessary optimality condition trivially holds. Thus, \tilde{x} is a strict local minimum, according to Theorem 6.37.

(b3) The analysis for the point $\hat{x} = (1, 0)^T$ is identical to the one made in (b2). We can thus conclude that \hat{x} is another strict local minimum. In fact, \hat{x} and \tilde{x} are both global minima.

References

1. Abadie, J.: On the Kuhn-Tucker theorem. In: Abadie, J. (ed.) Nonlinear Programming. North-Holland Publishing Co., Amsterdam (1967)
2. Alizadeh, F., Goldfarb, D.: Second-order cone programming. *Math. Program. Ser. B* **95**, 3–51 (2003)
3. Andersen, E.D., Roos, C., Terlaky, T.: On implementing a primal-dual interior-point method for conic quadratic optimization. *Math. Program. Ser. B* **95**, 249–277 (2003)
4. Anderson, E.J., Nash, P.: Linear Programming in Infinite-Dimensional Spaces: Theory and Applications. Wiley, Chichester (1987)
5. Andreescu, T., Mushkarov, O., Stoyanov, L.: Geometric Problems on Maxima and Minima. Birkhäuser Boston Inc, Boston (2006)
6. Anjos, M.F.: Conic linear optimization. In: [84], pp. 107–120
7. Anjos, M.F., Lasserre, J.-B. (eds.): Handbook on Semidefinite, Conic and Polynomial Optimization. International Series in Operations Research & Management Science, Springer, New York (2011)
8. Bazaraa, M.S., Shetty, C.M., Sherali, H.D.: Nonlinear Programming: Theory and Algorithms. Wiley, New York (1983)
9. Beck, A., Sabach, S.: Weiszfeld's method: old and new results. *J. Opt. Theor. Appl.* **164**, 1–40 (2015)
10. Ben-Tal, A., Nemirovski, A.: Lectures on Modern Convex Optimization. Analysis, Algorithms, and Engineering Applications. MPS/SIAM Series on Optimization. SIAM, Philadelphia (2001)
11. Bertsekas, D.P.: Nonlinear Programming. Athena Science, Belmont (1999)
12. Best, M.J.: Quadratic Programming with Computer Programs. CRC Press, Boca Raton (2017)
13. Björk, A.: Numerical Methods for Least Squares Problems. SIAM Publications, Philadelphia (1996)
14. Blum, E., Oettli, W.: Direct proof of the existence theorem for quadratic programming. *Oper. Res.* **20**, 165–167 (1972)
15. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge U. P, Cambridge (2004)
16. Brinkhuis, J., Tikhomirov, V.: Optimization: Insights and Applications. Princeton University Press, Princeton (2005)
17. Botsko, M.W.: A first derivative test for functions of several variables. *Am. Math. Mon.* **93**, 558–561 (1986)
18. Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. *SIAM Rev.* **60**, 223–312 (2018)
19. Byrne, C.L.: A First Course in Optimization. CRC Press, Boca Raton (2015)
20. Cánovas, M.J.: The Karush-Kuhn-Tucker Optimality Conditions (in Spanish). University Miguel Hernández, Class-Notes (2012)

21. Carroll, C.W.: The created response surface technique for optimizing nonlinear restrained systems. *Oper. Res.* **9**, 169–184 (1961)
22. Chu, Y.J.: Generalization of some fundamental theorems on linear inequalities [in Chinese]. *Acta Math. Sinica* **16**, 25–40 (1966)
23. Cipra, B.A.: The best of the 20th century: editors name top 10 algorithms. *SIAM News* **33** No. 4 (2000)
24. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-free Optimization. MOS-SIAM Series on Optimization. SIAM, Philadelphia (2009)
25. Contesse, L., Hirari-Urruty, J.-B., Penot, J.-P.: Least squares solutions of linear inequality systems: a pedestrian approach. *RAIRO-Oper. Res.* **51**, 567–575 (2017)
26. Cottle, R.W.: A Theorem of Fritz John in Mathematical Programming. RAND Corporation Memo, RM-3858-PR (1963)
27. Crouzeix, J.-P., Keraghel, A., Sosa, W.: Programación Matemática Diferenciable (Spanish). Instituto de Matemática y Ciencias Afines. IMCA, Lima (2011)
28. Crouzeix, J.-P., Ocaña, E., Sosa, W.: Análisis Convexo (Spanish). Monografías del Instituto de Matemática y Ciencias Afines. IMCA, Lima (2003)
29. Crouzeix, J.-P., Ocaña, E.: Ejercicios Resueltos de Programación Matemática Diferenciable (Spanish). Universidad Nacional de Ingeniería, Editorial Universitaria, Lima (2011)
30. Dantzig, G.: Linear programming. In: Lenstra, J.K. et al. (eds.) History of Mathematical Programming: A Collection of Personal Reminiscences. North-Holland (1991)
31. Dinh, N., Jeyakumar, V.: Farkas' lemma: three decades of generalizations for mathematical optimization. *TOP* **22**, 1–22 (2014)
32. Elliott, R.J., Kopp, P.E.: Mathematics of Financial Markets. Springer, Berlin (2005)
33. Faigle, U., Kern, W., Still, G.: Algorithmic Principles of Mathematical Programming. Kluwer Texts in Mathematical Sciences, vol. 24. Kluwer Academic, Dordrecht (2002)
34. Farkas, Gy.: Theorie der einfachen Ungleichungen. *J. Reine Angew. Math.* **124**, 1–27 (1901)
35. Figueiredo, M.A.T., Nowak, R.D., Wright, S.J.: Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. *IEEE J. Sel. Top. Signal. Process.* **10**, 586–597 (2007)
36. Frank, M., Wolfe, P.: An algorithm for quadratic programming. *Nav. Res. Logist. Q.* **3**, 95–110 (1956)
37. Franklin, J.: Mathematical methods of economics. *Am. Math. Mon.* **90**, 229–244 (1983)
38. Frisch, R.: The logarithmic potential method for solving linear programming problems. Memorandum. University Institute of Economics, Oslo (1955)
39. Gass, S.I., Assad, A.A.: An Annotated Timeline of Operations Research: An Informal History. International Series in Operations Research & Management Science, vol. 75. Kluwer Academic Publishers, Boston (2005)
40. Giannessi, F., Maugeri, A.: Variational Inequalities and Network Equilibrium Problems. Springer, New York (2013)
41. Giorgi, G., Kjeldsen, T.H. (eds.): Traces and Emergence of Nonlinear Programming. Springer, Basel (2014)
42. Goberna, M.A., Jornet, V., Puente, R.: Optimización Lineal: Teoría. Métodos y Modelos (Spanish). McGraw-Hill, New York (2004)
43. Goberna, M.A., López, M.A.: Linear Semi-Infinite Optimization. Wiley, Chichester (1998)
44. Goberna, M.A., López, M.A.: Recent contributions to linear semi-infinite optimization. *4OR-Q. J. Oper. Res.* **15**, 221–264 (2017). Updated at <https://doi.org/10.1007/s10479-018-2987-8>
45. Golub, G.H., Van Loan, C.F.: Matrix Computations. Johns Hopkins University Press, Baltimore (1996)
46. Gondzio, J.: Interior point methods 25 years later. *Eur. J. Oper. Res.* **218**, 587–601 (2012)
47. Grotschel, M. (ed.): Optimization Stories: 21st International Symposium on Mathematical Programming. Berlin, August 19–24, 2012. Doc. Math. Extra volume (2012)
48. Gueron, S., Tessler, R.: The Fermat-Steiner problem. *Am. Math. Mon.* **109**, 443–451 (2002)

49. Guignard, M.: Generalized Kuhn-Tucker conditions for mathematical programming problems in a Banach space. *SIAM J. Control* **7**, 232–241 (1969)
50. Hildebrandt, S., Tromba, A.: *The Parsimonious Universe. Shape and Form in the Natural World*. Copernicus, New York (1996)
51. Himmelblau, D.M.: *Applied Nonlinear Programming*. McGraw-Hill, New York (1972)
52. Hiriart-Urruty, J.-B.: *Optimisation et Analyse Convexe* (French). Presses Universitaires de France, Paris (1998)
53. Hiriart-Urruty, J.-B.: *Les Mathématiques du Mieux Faire* (French). Opuscules de l'Université Paul Sabatier (Toulouse III). Ellipses, Paris (2008)
54. Hiriart-Urruty, J.-B.: *Mathematical Tapas*. Springer Undergraduate Mathematics Series, vol. 1 (for Undergraduates). Springer, Cham (2016)
55. Hiriart-Urruty, J.-B., Laurent, P.-J.: A characterization by optimization of the orthocenter of a triangle. *Elemente Der Mathematik* **70**, 45–48 (2015)
56. Hiriart-Urruty, J.-B., Lemaréchal, C.: *Fundamentals of Convex Analysis: Abridged Version of Convex Analysis and Minimization Algorithms I*. Springer, Berlin (2001)
57. Hiriart-Urruty, J.-B., Malick, J.: A fresh variational-analysis look at the positive semidefinite matrices world. *J. Optim. Theory Appl.* **153**, 551–577 (2012)
58. Horn, R.A.: *Johnson, Ch.R: Matrix Analysis*. Cambridge U.P, New York (1990)
59. Karmarkar, N.: A new polynomial-time algorithm for linear programming. *Combinatorica* **4**, 373–395 (1984)
60. Kuhn, H.W., Tucker, A.W.: Nonlinear programming. In: Neyman, J. (ed.) *Proceedings of the 2nd Berkeley Symposium on Mathematical Statistics and Probability*. University of California Press, Berkeley (1951)
61. Kuttler, K.: *Calculus. Applications and Theory*. World Scientific, Singapore (2008)
62. Lenstra, J.K., Rinnooy Kan, A.H.G., Schrijver, A. (eds.): *History of Mathematical Programming: A Collection of Personal Reminiscences*. CWI, North-Holland Publishing Co., Amsterdam (1991)
63. Mangasarian, O.: Generalized support vector machines. In: *Advances in Large Margin Classifiers*, pp. 135–146. MIT Press, Cambridge (2000)
64. Mangasarian, O.L., Fromovitz, S.: The Fritz John necessary optimality conditions in the presence of equality and inequality constraints. *J. Math. Anal. Appl.* **17**, 37–47 (1967)
65. McKinnon, K.I.M.: Convergence of the Nelder-Mead Simplex Method to a nonstationary point. *SIAM J. Optim.* **9**, 148–158 (1998)
66. Monteiro, R.D.C., Adler, I.: Interior path following primal-dual algorithms. Part I: linear programming. *Math. Program.* **44**, 27–41 (1989)
67. Mordukhovich, B.S.: *Variational Analysis and Applications*. Springer, New York (2018)
68. Nahin, P.J.: *When Least Is Best: How Mathematicians Discovered Many Clever Ways to Make Things as Small (or as Large) as Possible*. Princeton University Press, Princeton (2004)
69. Nesterov, Y.E., Todd, M.J.: Primal-dual interior-point methods for self-scaled cones. *SIAM J. Optim.* **8**, 324–364 (1998)
70. Nocedal, J., Wright, S.J.: *Numerical Optimization*, 2nd edn. Springer, New York (2006)
71. Oldfather, W.A., Ellis, C.A., Brown, D.M.: Leonhard Euler's elastic curves. *Isis* **20**, 72–160 (1933)
72. Oliveira, O.R.B.: The fundamental theorem of algebra: from the four basic operations. *Am. Math. Mon.* **119**, 753–775 (2012)
73. Osborne, M.R., Presnell, B., Turlach, B.A.: A new approach to variable selection in least squares problems. *IMA J. Numer. Anal.* **20**, 389–403 (2000)
74. Pataki, G.: Strong duality in conic linear programming: facial reduction and extended duals. In: Bailey, D. et al. (eds) *Computational and Analytical Mathematics*. Springer Proceedings in Mathematics & Statistics, vol. 50. Springer, New York (2013)
75. Peressini, A.L., Sullivan, F.E., Uhl, J.J.: *The Mathematics of Nonlinear Programming*. Springer, Berlin (1988)
76. Peterson, D.W.: A review of constraint qualifications in finite-dimensional spaces. *SIAM Rev.* **15**, 639–654 (1973)

77. Powell, M.J.D.: On search directions for minimization algorithms. *Math. Program.* **4**, 193–201 (1973)
78. Prékopa, A.: On the development of optimization theory. *Am. Math. Mon.* **87**, 527–542 (1980)
79. Robbins, H., Monro, S.: A stochastic approximation method. *Ann. Math. Stat.* **22**, 400–407 (1951)
80. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1970)
81. Rosenbrock, H.H.: An automatic method for finding the greatest or least value of a function. *Comput. J.* **3**, 175–184 (1960)
82. Sørensen, K.: Metaheuristics - the metaphor exposed. *Int. Trans. Oper. Res.* **22**, 3–18 (2015)
83. Stoer, J., Bulirsch, R.: Introduction to Numerical Analysis, 3rd edn. Springer, New York (2002)
84. Tatsumi, K., Tanino, T.: Support vector machines maximizing geometric margins for multi-class classification. *TOP* **22**, 856–859 (2014)
85. Terlaky, T., Anjos, M.F., Ahmed, S. (eds.): Advances and Trends in Optimization with Engineering Applications. MOS-SIAM Series on Optimization, vol. 24. SIAM, Philadelphia (2017)
86. Terkelsen, F.: The fundamental theorem of algebra. *Am. Math. Mon.* **83**, 647 (1976)
87. Turlach, B.A., Wright, S.J.: Quadratic programming. *Wiley Interdiscip. Rev. Comput. Stat.* **7**, 153–159 (2015)
88. Van der Vorst, H.A.: Iterative Krylov Methods for Large Linear Systems. Cambridge U. P, Cambridge (2003)
89. Watkins, D.S.: The Matrix Eigenvalue Problem: GR and Krylov Subspace Methods. SIAM, Philadelphia (2007)
90. Wolfe, P.: A duality theorem for non-linear programming. *Q. Appl. Math.* **19**, 239–244 (1961)

Index

A

Abadie Constraint Qualification (ACQ), 283
Active index, 134, 276
Armijo rule, 188
Asymptotic notation, 23

B

Big data, 102
Broyden–Fletcher–Goldfarb–Shanno (BFGS), 229

C

Characteristic polynomial, 29
Complementarity Condition (CC), 138, 279
Computational complexity, 197
Condition
 complementarity (CC), 138, 279
 curvature, 189
 Goldstein, 192
 Karush–Kuhn–Tucker (KKT), 138, 164, 279
 nonnegativity (NC), 138
 saddle point (SPC), 162
 stationarity (SC), 138
 Wolfe, 189
 Zoutendijk, 195

Condition number, 31

Cone, 58
 active, 135, 276
 closed convex, 61
 feasible directions, 134, 135
 of positive semidefinite symmetric matrices, 172
 pointed, 173
 polar, 135

second order, 172
symmetric, 173
tangent, 278

Constraint
 nondegenerate, 293
 set, 7
 strongly active, 292
Constraint Qualification (CQ)
 Abadie (ACQ), 283
 Guignard (GCQ), 280, 283
 Kuhn–Tucker (KTCQ), 283
 linear independence (LICQ), 284
 Mangasarian–Fromovitz (MFCQ), 283
 Slater (SCQ), 156, 284

Convergence
 geometric, 199
 global, 193
 linear, 198
 local, 194, 211
 quadratic, 198
 sublinear, 198
 superlinear, 198
 superlinear of order p , 198

Cosine measure, 238

D

Derivative, 18, 62
 directional, 22
 left and right, 62
 partial, 18
Differential of a function, 18, 22
Direction
 descent, 184
 feasible, 134
 Newton, 186, 209
 Q -conjugate, 219

- search, 184
- D**
- Domain
of a function, 17
of a multifunction, 154
- Duality gap, 168
- Dual problem
conic, 174
geometric, 111
Haar, 173
Lagrange, 168
penalty, 257
Wolfe, 170, 171
- E**
- Epigraph
of a function, 72
of a multifunction, 154
- Estimator
LASSO, 146
maximum likelihood, 134
- F**
- Farkas lemma, 136
- Fermat principle, 33
- Fermat–Steiner problem, 123, 129
- Function
affine, 9
augmented Lagrangian, 274
barrier, 261
coercive, 39, 40, 86
concave, 62, 155
constraint, 7
convex, 62, 70–73, 77, 79, 80, 104, 155
differentiable, 22, 62
error, 197
extended real, 154
increasing, 10
Lagrangian, 161
linear, 9
Lipschitz continuous, 68, 77
multivariate, 25
objective, 7
of several variables, 17
penalty, 255
posynomial, 9
quadratic, 9, 92
section, 76, 77
strictly convex, 82
strongly convex, 83–86
value, 154, 156
- G**
- Gradient, 20
monotone, 79
strictly monotone, 83
- Graph of a function, 17
- Guignard Constraint Qualification (GCQ), 280, 283
- H**
- Hull
conical, 58, 59
convex, 55
- Hyperplane
nonvertical support, 74
strict separation, 56
support, 57, 58, 61
- I**
- Inequality
arithmetic-quadratic, 105
first of Jensen, 104
geometric-arithmetic, 107
geometric-quadratic, 106
Kantorovich, 203
second of Jensen, 105
- K**
- Karush–Kuhn–Tucker (KKT), 138, 164, 279
KKT system, 140
Kuhn–Tucker Constraint Qualification (KTCQ), 283
- L**
- Lagrange multipliers, 267
Landau notation, 23
Least Absolute Shrinkage and Selection Operator (LASSO), 146
Level curve, 17
Linear Independence Constraint Qualification (LICQ), 284
Line search, 184
Lipschitz continuity, 68, 77
- M**
- Machine learning, 102, 143
Mangasarian–Fromovitz Constraint Qualification (MFCQ), 283
- Matrix
gradient, 26

- Gram, 27
 Hessian, 23
 ill-conditioned, 32
 indefinite, 27, 28
 invariants, 30
 Jacobian, 26
 negative definite, 27
 negative semidefinite, 27
 positive definite, 27, 28, 30
 positive semidefinite, 27, 28, 30
 spectral decomposition, 28
 well-conditioned, 32
- Method
 backtracking, 192
 barrier, 263
 Broyden–Fletcher–Goldfarb–Shanno (BFGS), 229
 conjugate directions, 220
 conjugate gradient, 222
 coordinate descent, 233
 Davidon–Fletcher–Powell (DFP), 228
 derivative-free, 232
 directional direct-search, 236
 double sweep of Aitken, 233
 exterior penalty, 257
 Fourier elimination, 59
 Gauss–Newton, 214
 homotopy, 146
 interior point, 260
 Lagrange multipliers, 267
 Levenberg–Marquardt, 215
 Newton, 186, 209, 216
 penalty, 256
 quasi-Newton, 226
 sequential quadratic programming (SQP), 297
 simplex of Nelder and Mead, 234
 steepest descent, 185, 202
 trust region, 212
 Weiszfeld, 130
- Metric projection, 148, 151
- Minimum
 global, 8, 38, 40, 73
 local, 8, 33, 35
 nonsingular, 37
 strict, 9, 33, 35
- Monomial, 109
- Multifunction, 147
 feasible set, 154
- Multivariate chain rule, 26
- O**
- Optimality conditions
 first order, 33
 Fritz John, 288
 Lagrange, 267, 275
 second order, 35, 293–295
- Optimal value, 8
- Optimization
 conic, 118
 convex, 117
 geometric, 9, 109, 111
 linear, 9, 138
 nonlinear, 9
 quadratic, 9, 93, 141, 147, 148
- P**
- Path of minimizers, 146
- Point
 critical, 33, 35
 KKT, 279
 saddle, 33, 35, 162, 166
 Slater, 156, 260
 stationary, 33
- Polytope, 55
- Positive span, 237
- Posynomial, 109
- Principal minor, 30
 director, 30
- Problem
 bounded, 8
 inconsistent, 8
 least-squares, 214
 linear conic, 173
 linear semi-infinite, 173
 parametric, 153
 relaxed, 140
 solvable, 8, 38
 unbounded, 8
 unconstrained, 7
- R**
- Regression
 least squares line, 95
 linear, 14, 16, 95
 polynomial, 98
 with interpolation, 151
- Relative interior, 174
- S**
- Nonnegativity Condition (NC), 138
- Saddle Point Condition (SPC), 162

- Section of a function, 76, 77
 Sensitivity analysis, 117, 158, 166, 296
 Sequential Quadratic Programming (SQP), 297
- Set
 closed convex, 56, 58
 convex, 55, 56
 feasible, 8
 optimal, 8, 93, 118
 polar, 137
 sublevel, 38
- Set-valued mapping, 147
- Slater Constraint Qualification (SCQ), 156, 284
- Solution
 approximate, 102, 119
 feasible, 7
 optimal, 9, 102
- Spectral
 decomposition, 28
 norm, 31
 radius, 31
- Stationarity Condition (SC), 138
- Stepsize, 184
- Strict separation, 56, 61
- Subgradient, 160
- Sufficient decrease, 188
- T**
- Tangent
 left-hand, 62
 plane, 20
 right-hand, 62
- Term
- regularization, 145
 sparsity, 145
- Theorem
 barrier methods, 263
 convergence of directional direct-search methods, 239
 convergence of Newton's method, 209, 217
 convergence of penalty methods, 258
 convergence of SQP methods, 299
 Frank-Wolfe, 141
 Fritz John optimality conditions, 288
 fundamental of algebra, 46
 global optimality conditions, 288
 Gordan, 284
 KKT optimality conditions, 280, 285
 KKT with convex constraints, 164
 KKT with linear constraints, 138
 Stoltz, 63
 strong geometric duality, 111
 strong Lagrange duality, 168
 strong Wolfe duality, 171
 Weierstrass, 34
 Zoutendijk, 193
- Trust region, 212
- V**
- Vector
 KKT, 139
 Lagrange, 163
 residual, 14, 95, 98
 right-hand side, 154
 sensitivity, 153, 160
 sparse, 145