

Boris T. Polyak

INTRODUCTION  
TO  
OPTIMIZATION



## TRANSLATIONS SERIES IN MATHEMATICS AND ENGINEERING

---

A.V. Balakrishnan  
General Editor

*This is the revised version of the book, originally published in 1987. All corrections are made with proof-reading marks on the margins.*

*I am indebted to numerous readers of the monograph who indicated typos and inaccuracies in the original text. The contribution of my friend Olvi Mangasarian and his students was extraordinary helpful.*

*My colleague Andrey Tremba incorporated all revisions in the text; I highly appreciate his assistance.*

Boris Polyak

November 2010.

# TRANSLATIONS SERIES IN MATHEMATICS AND ENGINEERING

---

M.I. Yadrenko  
**Spectral Theory of Random Fields**  
1983, 267 pp.  
ISBN 0-911575-00-6

G.I. Marchuk  
**Mathematical Models in Immunology**  
1983, 378 pp.  
ISBN 0-911575-01-4

A.A. Borovkov, ed.  
**Advances In Probability Theory:  
Limit Theorems and Related Problems**  
1984, 392 pp.  
ISBN 0-911575-03-0

V.A. Dubovitskij  
**The Ulam Problem of Optimal Motion  
of Line Segments**  
1985, 128 pp.  
ISBN 0-911575-04-9

N.V. Krylov, R.S. Liptser, and  
A.A. Novikov, eds.  
**Statistics and Control of  
Stochastic Processes**  
1985, 521 pp.  
ISBN 0-911575-18-9

Yu.G. Evtushenko  
**Numerical Optimization Techniques**  
1985, 575 pp.  
ISBN 0-911575-07-3

V.F. Dem'yanov, and L.V. Vasil'ev  
**Nondifferentiable Optimization**  
1985, 472 pp.  
ISBN 0-911575-09-X

A.A. Borovkov, ed.  
**Advances In Probability Theory:  
Limit Theorems for Sums of Random  
Variables**  
1985, 313 pp.  
ISBN 0-911575-17-0

V.F. Kolchin  
**Random Mappings**  
1986, 224 pp.  
ISBN 0-911575-16-2

L. Telksnys, ed.  
**Detection of Changes in Random  
Processes**  
1986, 240 pp.  
ISBN 0-911575-20-0

V.F. Dem'yanov, and A.M. Rubinov  
**Quasidifferential Calculus**  
1986, 301 pp.  
ISBN 0-911575-35-9

V.P. Chistyakov, B.A. Sevast'yanov,  
and V.K. Zakharov  
**Probability Theory for Engineers**  
1987, 175 pp.  
ISBN 0-911575-13-8

B.T. Polyak  
**Introduction to Optimization**  
1987, 464 pp.  
ISBN 0-911575-16-6

---

A.V. Balakrishnan, A.A. Dorodnitsyn,  
and J.L. Lions, eds.  
**Vistas in Applied Mathematics:  
Numerical Analysis, Atmospheric  
Sciences, Immunology.**  
1986, 396 pp.  
ISBN 0-911575-38-3

R. Kalman, A. Viterbi, et al.  
**Recent Advances in Communication  
and Control Theory**  
1987, approx. 450 pp.  
ISBN 0-911575-46-4

74

BORIS T. POLYAK

INTRODUCTION  
TO  
OPTIMIZATION



Optimization Software, Inc.  
Publications Division, New York

*Author*

B.T. Polyak  
Institute of Control Science  
65 Profsoyuznaya ulitsa  
Moscow 117342  
U.S.S.R.

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

Poliak, B.T. (Boris Teodorovich)  
Introduction to optimization.

(Translations series in mathematics and engineering)

Translation of: Vvedenie v optimizatsiiu.

Bibliography: p.

Includes index.

1. Mathematical optimization. I. Title. II. Series.

QA402.5.P58313 1987 519 87-11290

ISBN 0-911575-14-6

© 1987 by Optimization Software, Inc., Publications Division,  
4 Park Avenue, New York, New York 10016. All rights reserved.  
Published in 1987. Printed in the United States of America.

## **ABOUT THE AUTHOR**

Boris Teodorovich Polyak was born in Moscow in 1935. He received his *Candidat* degree in Physical-Mathematics in 1964 from the Moscow State University. From 1964 to 1971 he was with the Computer Center there. Since 1971 he has been with the Institute of Control Sciences, Moscow, where he is currently a Senior Scientist. He holds the degree of Doctor of Engineering Sciences.

His main research interests include: Numerical Analysis, Optimization Theory, Mathematical Programming, and Recursive Estimation, and he is the author of over 80 published papers. He is also on the Editorial Board of *Numerical Functional Analysis & Optimization* and *Automation & Remote Control*.



## TABLE OF CONTENTS

<b>Foreword</b>	xv
<b>Preface</b>	xvii
<b>Introduction</b>	xix
<b>Notation</b>	xxv
<b>Part I. UNCONSTRAINED MINIMIZATION</b>	1
<b>Chapter 1. Fundamentals of the Theory and Methods of Unconstrained Minimization</b>	2
<b>1.1 REVIEW OF MATHEMATICAL ANALYSIS</b>	2
1.1.1 Differentiation of Scalar Functions	2
1.1.2 Differentiation of Vector Functions	5
1.1.3 Second Derivatives	6
1.1.4 Convex Functions	8
<b>1.2 EXTREMUM CONDITIONS</b>	11
1.2.1 A First-order Necessary Condition	11
1.2.2 A First-order Sufficient Condition	12
1.2.3 A Second-order Necessary Condition	12
1.2.4 A Second-order Sufficient Condition	13
1.2.5 What are Extremum Conditions Good for?	14
<b>1.3 EXISTENCE, UNIQUENESS, AND STABILITY OF A MINIMUM</b>	14
1.3.1 Existence of a Minimum	14
1.3.2 Uniqueness of a Solution	15
1.3.3 Stability of a Solution	16
<b>1.4 THE GRADIENT METHOD</b>	20
1.4.1 Heuristic Considerations	20
1.4.2 Convergence	21
<b>1.5 NEWTON'S METHOD</b>	27
1.5.1 Heuristic Considerations	27
1.5.2 Convergence	28
1.5.3 Newton's Method for Solving Equations	31
<b>1.6 THE ROLE OF CONVERGENCE THEOREMS</b>	31
1.6.1 Extreme Viewpoints	31
1.6.2 Why are Convergence Theorems Necessary?	32
1.6.3 Proceed with Caution	34

<b>Chapter 2. General Schemes for Investigating Iterative Methods</b>	<b>37</b>
<b>2.1 LYAPUNOV'S FIRST METHOD</b>	<b>37</b>
2.1.1 Review of Linear Algebra	37
2.1.2 Theorems on Linear Convergence	40
2.1.3 A Theorem on Superlinear Convergence	42
<b>2.2 LYAPUNOV'S SECOND METHOD</b>	<b>43</b>
2.2.1 Lemmas on Numerical Sequences	43
2.2.2 Lemmas on Random Sequences	47
2.2.3 The Main Theorems	50
2.2.4 Possible Modifications	54
<b>2.3 OTHER SCHEMES</b>	<b>56</b>
2.3.1 The Contraction Mapping Principle	56
2.3.2 The Implicit Function Theorem	57
2.3.3 The Role of General Schemes for Investigating Convergence	58
<b>Chapter 3. Minimization Methods</b>	<b>59</b>
<b>3.1 MODIFICATIONS OF THE GRADIENT METHOD AND OF NEWTON'S METHOD</b>	<b>59</b>
3.1.1 Advantages and Drawbacks of the Earlier Methods	59
3.1.2 Modifications of the Gradient Method	60
3.1.3 Modifications of Newton's Method	63
<b>3.2 MULTISTEP METHODS</b>	<b>65</b>
3.2.1 The Heavy Ball Method	65
3.2.2 The Conjugate Gradient Method	68
<b>3.3 OTHER FIRST ORDER METHODS</b>	<b>75</b>
3.3.1 Quasi-Newton Methods	75
3.3.2 Methods of Variable Metric and Methods of Conjugate Directions	78
3.3.3 The Secant Method	81
3.3.4 Other Approaches for Constructing the First-order Methods	83
<b>3.4 DIRECT METHODS</b>	<b>87</b>
3.4.1 General Characteristics	87
3.4.2 Methods of Linear Approximation	87
3.4.3 Nonlocal Linear Approximation	90
3.4.4 Quadratic Approximation	92

<b>Chapter 4. Influence of Noise</b>	95
<b>4.1 SOURCES AND TYPES OF NOISE</b>	95
4.1.1 Sources of Noise	95
4.1.2 Types of Noise	97
<b>4.2 THE GRADIENT METHOD IN THE PRESENCE OF NOISE</b>	98
4.2.1 The Statement of the Problem	98
4.2.2 Absolute Deterministic Noise	98
4.2.3 Relative Deterministic Noise	100
4.2.4 Absolute Random Noise	100
4.2.5 Relative Random Noise	102
<b>4.3 OTHER MINIMIZATION METHODS IN THE PRESENCE OF NOISE</b>	103
4.3.1 Newton's Method	103
4.3.2 Multistep Methods	104
4.3.3 Other Methods	105
<b>4.4 DIRECT METHODS</b>	106
4.4.1 The Statement of the Problem	106
4.4.2 Difference Methods for Random Noise	106
4.4.3 Other Methods	109
<b>4.5 OPTIMAL METHODS IN THE PRESENCE OF NOISE</b>	111
4.5.1 Potential Possibilities of Iterative Methods in the Presence of Noise	111
4.5.2 Optimal Algorithms	116
<b>Chapter 5. Minimization of Nondifferentiable Functions</b>	119
<b>5.1 CONVEX ANALYSIS: FUNDAMENTALS</b>	119
5.1.1 Convex Sets and Projection	120
5.1.2 Separation Theorems	122
5.1.3 Convex Nondifferentiable Functions	124
5.1.4 The Subgradient	127
5.1.5 The $\varepsilon$ -subgradient	132
<b>5.2 EXTREMUM CONDITIONS, EXISTENCE, UNIQUENESS,         AND STABILITY OF A SOLUTION</b>	133
5.2.1 Extremum Conditions	133
5.2.2 Existence and Uniqueness of a Minimum	135
5.2.3 Stability of a Minimum	135

<b>5.3 THE SUBGRADIENT METHOD</b>	138
5.3.1 The Substance of the Method	138
5.3.2 The Main Results	140
5.3.3 The $\varepsilon$ -subgradient Method	144
<b>5.4 ALTERNATIVE METHODS</b>	145
5.4.1 Preliminary Remarks	145
5.4.2 Multistep Methods	146
5.4.3 Optimal Methods	153
5.4.4 Space Extension Methods	154
<b>5.5 THE INFLUENCE OF NOISE</b>	158
5.5.1 The Statement of the Problem	158
5.5.2 Absolute Deterministic Noise	158
5.5.3 Relative Deterministic Noise	159
5.5.4 Absolute Random Noise	159
<b>5.6 SEARCH METHODS</b>	160
5.6.1 The One-dimensional Case	160
5.6.2 The Multidimensional Case	162
<b>Chapter 6. Singularity, Multimodality, Nonstationarity</b>	165
<b>6.1 A SINGULAR MINIMUM</b>	165
6.1.1 The Behavior of Standard Methods	165
6.1.2 Special Methods for Singular Problems	173
6.1.3 Methods in the Presence of Noise	179
6.1.4 Summary	182
<b>6.2 MULTIMODALITY</b>	185
6.2.1 Preliminary Remarks	185
6.2.2 Exact Methods	187
6.2.3 Deterministic Heuristic Methods	189
6.2.4 Stochastic Heuristic Methods	192
<b>6.3 NONSTATIONARY PROBLEMS</b>	194
6.3.1 The Form of $f(x, t)$ is Known	194
6.3.2 The Form of $f(x, t)$ is Unknown	195
6.3.3 Summary	196

<b>Part II. CONSTRAINED MINIMIZATION</b>	<b>199</b>
<b>Chapter 7. Minimization on Simple Sets</b>	<b>200</b>
<b>7.1 THEORETICAL FOUNDATIONS</b>	<b>200</b>
7.1.1 Extremum Conditions in the Smooth Case	200
7.1.2 Extremum Conditions in the Convex Case	203
7.1.3 Existence, Uniqueness and Stability of a Minimum	204
<b>7.2 BASIC METHODS</b>	<b>206</b>
7.2.1 The Gradient Projection Method	206
7.2.2 The Subgradient Projection Method	210
7.2.3 The Conditional Gradient Method	210
7.2.4 Newton's Method	214
<b>7.3 OTHER METHODS</b>	<b>216</b>
7.3.1 Quasi-Newton Methods	216
7.3.2 The Conjugate Gradient Method	219
7.3.3 Minimization of Nonsmooth Functions	221
<b>7.4 THE INFLUENCE OF NOISE</b>	<b>221</b>
7.4.1 Absolute Deterministic Noise	221
7.4.2 Absolute Random Noise	222
7.4.3 Relative Noise	223
<b>Chapter 8. Problems with Equality Constraints</b>	<b>224</b>
<b>8.1 THEORETICAL FOUNDATIONS</b>	<b>224</b>
8.1.1 Lagrange Multipliers	224
8.1.2 Second-order Minimum Conditions	230
8.1.3 The Usage of Extremum Conditions	233
8.1.4 Existence, Uniqueness and Stability of a Solution	234
<b>8.2 MINIMIZATION METHODS</b>	<b>237</b>
8.2.1 Classification of the Methods	237
8.2.2 The Linearization Method	238
8.2.3 Dual Methods	240
8.2.4 The Augmented Lagrangian Method	241
8.2.5 The Penalty Function Method	244
8.2.6 The Reduced Gradient Method	245
8.2.7 Newton's Method	246
8.2.8 Other Quadratically Convergent Methods	247

8.3 HOW TO HANDLE POSSIBLE COMPLICATIONS	248
8.3.1 A Global Minimum	248
8.3.2 Noise	249
8.3.3 A Singular Minimum	251
8.3.4 Incompatibility of Constraints	252
<b>Chapter 9. The General Problem of Mathematical Programming</b>	<b>253</b>
9.1 THE THEORY OF CONVEX PROGRAMMING	253
9.1.1 Convex Analysis: Fundamentals	253
9.1.2 The Kuhn-Tucker Theorem	259
9.1.3 Duality	265
9.1.4 Existence, Uniqueness and Stability of a Solution	268
9.2 NONLINEAR PROGRAMMING (THEORY)	270
9.2.1 Necessary Conditions for a Minimum	270
9.2.2 Sufficient Conditions for a Minimum	274
9.2.3 Uniqueness and Stability of a Solution	276
9.3 CONVEX PROGRAMMING METHODS	279
9.3.1 Methods of Feasible Directions	280
9.3.2 The Linearization Method	282
9.3.3 Dual Methods	283
9.3.4 Penalty Methods and Related Methods	288
9.3.5 Methods for Nonsmooth Problems	292
9.3.6 Summary	296
9.4 NONLINEAR PROGRAMMING METHODS	296
9.4.1 The Linearization Method	297
9.4.2 Newton-like and Quasi-Newton Methods	298
9.4.3 Other Methods	300
<b>Chapter 10. Linear and Quadratic Programming</b>	<b>302</b>
10.1 LINEAR PROGRAMMING (THEORY)	302
10.1.1 Types of Problems	302
10.1.2 Structure of Polyhedral Sets	304
10.1.3 Extremum Conditions	309
10.1.4 Existence, Uniqueness and Stability of a Solution	312
10.2 FINITE LINEAR PROGRAMMING METHODS	317
10.2.1 The Simplex Method	317
10.2.2 Implementation of the Simplex Method	320
10.2.3 Other Finite Methods	321
10.2.4 Why Does the Simplex Method Work?	323

<b>10.3 ITERATIVE METHODS OF LINEAR PROGRAMMING</b>	<b>325</b>
10.3.1 The Need for Iterative Methods	325
10.3.2 Iterative Finite Methods	326
10.3.3 Reduction to Nonsmooth Minimization	329
10.3.4 The Lagrange Functions	332
10.3.5 Summary	334
<b>10.4 QUADRATIC PROGRAMMING</b>	<b>334</b>
10.4.1 Extremum Conditions	335
10.4.2 Existence, Uniqueness and Stability of a Solution	337
10.4.3 Finite Methods	338
10.4.4 Iterative Methods	339
<b>Chapter 11. Optimization Problems: Examples</b>	<b>342</b>
<b>11.1 IDENTIFICATION PROBLEMS</b>	<b>342</b>
11.1.1 Statistical Problems of Parameter Estimation	343
11.1.2 Regression Problems	345
11.1.3 Robust Estimation	347
11.1.4 Recursive Estimation	350
11.1.5 Data Analysis	352
11.1.6 Other Identification Problems	356
<b>11.2 OPTIMIZATION PROBLEMS IN ENGINEERING AND ECONOMICS</b>	<b>358</b>
11.2.1 Optimal Design	358
11.2.2 Optimal Allocation of Resources	360
11.2.3 Optimal Planning	361
11.2.4 Optimization under Uncertainty	364
11.2.5 Extremal Control	366
11.2.6 Optimal Control	367
<b>11.3 OPTIMIZATION PROBLEMS IN MATHEMATICS AND PHYSICS</b>	<b>371</b>
11.3.1 Optimal Approximation Problems	371
11.3.2 Geometric Extremum Problems	373
11.3.3 Variational Principles in Physics	375

<b>Chapter 12. Optimization Problems: Implementation</b>	<b>377</b>
<b>12.1 SOLUTION OF A PROBLEM</b>	<b>377</b>
12.1.1 The Mathematical “Formalization” of a Problem	377
12.1.2 The Choice of Methods and Codes	379
12.1.3 Evaluation of Solutions	380
<b>12.2 OPTIMIZATION SOFTWARE</b>	<b>381</b>
12.2.1 General Requirements	381
<b>12.3 Test Problems and Computational Results</b>	<b>381</b>
12.3.1 Criteria for a Comparative Analysis of Algorithms. Empirical Results	382
12.3.2 Test Problems: General Requirements	383
12.3.3 Unconstrained Minimization of Smooth Functions	384
12.3.4 Unconstrained Minimization of Nonsmooth Functions	391
12.3.5 Nonlinear Programming	393
12.3.6 Linear Programming	398
<b>Notes</b>	<b>401</b>
<b>References</b>	<b>415</b>
<b>Index</b>	<b>434</b>

## **FOREWORD**

The field of nonlinear optimization has benefited from several important ideas developed in the Soviet Union. Some of these ideas underwent a parallel development in the West, but others received inadequate attention in English language textbooks. For this reason the publication of the present book by a principal Soviet contributor is particularly valuable. It represents what is probably the first comprehensive synthesis of the nonlinear programming methodologies that are popular in the West and the Soviet Union.

The reader will find here a systematic treatment of both classical subjects, and topics little covered elsewhere—such as nondifferentiable optimization, degenerate problems, and stochastic optimization methods. Beyond this, however, this text has many significant merits. It gives careful attention to both mathematical rigor and practical relevance. The convergence analysis of numerical methods is done in a unified manner. A systematic effort is made to chart the limits of the methodology by providing performance analysis on difficult problems. There is a thoughtful discussion of the practical solution process. A wealth of new or little known material is included in the text and the exercises. Above all, the book is written by a true expert with a refined understanding of the nature, purpose, and limitations of nonlinear optimization and applied mathematics in general.

Dimitri P. Bertsekas  
Professor of Electrical Engineering  
Massachusetts Institute of Technology



## PREFACE

The extraordinary ubiquity of optimization problems in engineering, economics and management has rendered necessary that a broad group of practitioners be familiar with methods for solving such problems. It is however difficult for an engineer or economist to orient herself/himself in the enormous literature on optimization—most of the existent published works have been written “by mathematicians for mathematicians”—and work his or her way through the maze of problems and algorithms. In this book, we endeavor to present systematically the current theory and methods of optimization in the form comprehensible to the engineer. Only a minimum of mathematical prerequisites is required: the basics of Mathematical Analysis, Linear Algebra and Probability Theory are sufficient. As the exposition progresses, the problems become more complicated. We begin with the simplest problems of unconstrained minimization of smooth functions and proceed to investigate the influence of different complicating factors—such as noise, nonsmooth functions, singularity of a minimum, and constraints. Problems of each class are analyzed in a similar way: first we develop the necessary mathematics, next prove conditions for an extremum, followed by results on existence, uniqueness and stability of a solution, and finally the numerical methods for solving the problems. We focus our attention on the general notion on which the methods would be based, present a comparative analysis, demonstrating how the theoretical results provide the foundation for the methods developed. We illustrate the relationship between the general and the particular methods using particular optimization problems as examples. An extensive list of references for further study is also provided.

The material in this book is rather different from the traditional. Textbooks in Mathematical Programming treat—more or less exclusively—the simplex method of linear programming. We limit our treatment of it to one brief section. On the other hand, we do pay a great deal of attention to the problem of unconstrained minimization which serves as a vehicle for discussing the basic concepts of the theory as well as the methods of optimization. Some of the non-conventional interpretations are those of nonsmooth optimization problems, singular and nonstationary problems, equality-constraint problems, stability conditions for an extremum, effect of noise on optimization methods, analysis of general schemes of investigating the convergence of iterative methods, among others. We systematically discuss some “naive” questions, not usually addressed in the mathematical literature, e.g., What do you need extremum conditions for? What can be gained from theoretical results on the convergence of methods? Can unstable optimization problems really be solved?

The book deals only with finite-dimensional problems. The reason is twofold: space limitation on the book and the prerequisite background in mathematics. We have omitted discussing optimality conditions in general extremum problems, problems of variational calculus and optimal control problems, and some others. Also, the ideas and results generated in the finite-dimensional problems provide the models for more general optimization problems. The reader conversant with Functional Analysis should have no difficulty in recognizing that many assertions go over to Hilbert and/or Banach spaces—however, no such generalizations are given in the book. We also skipped discrete optimization problems, because they call for entirely different methods of investigation, relying more on Combinatorics and Mathematical Logic.

The author has lectured on the theory and techniques of optimization, for example, at the Moscow State University and the Institute of Control Sciences. These occasions have provided ample evidence for the differences in approach between mathematicians, computer analysts and users. This book is an attempt to find a compromise solution, to meet the needs of this diverse audience. Addressing mathematicians, the author wishes to point out that this is not a textbook on “optimization methods”—some theorems have not been proved and a lot of material is given in the form of exercises for the reader to work on her/his own. Nor will the computing analyst find the determinate formulations of algorithms or ready-to-use computer software. Some results are only of theoretical interest—in other words, this book is not a collection of recipes to solve specific problems. The third group of users—engineers and economists, the author hopes, will bear with an often abstract mode of presentation: examples and applications are given only in the concluding chapters.

The idea of writing this book originates with Ya.Z. Tsyplkin, whose vast knowledge and significant contributions in the area of optimization problems have been of great value to the author during many years of our cooperation. The computational expertise of E.N. Belov and B.A. Skokov has been an essential contribution. Yu.E. Nesterov rendered invaluable assistance in editing the text, G.M. Korpelevich in improving the presentation; and G.N. Arkhipova in the preparation of the manuscript. To all of them the author wishes to express his sincere gratitude.

## INTRODUCTION

As a rule, when many options are available, man's actions are guided by the need to choose the best possible way. Human activity, indeed, implicates solving (consciously or unconsciously) optimization problems. Moreover, many laws of nature are of a variational character, even if inappropriate in this case to speak of the existence of a purpose.

One might think that this omnipresence of optimization problems would be reflected in mathematics. But the fact is that mathematicians have been tackling extremum problems only sporadically over many centuries, and the theory and techniques for solving such problems started to burgeon only as recently as the 1950s.

The elementary problem of unconstrained minimization of a function of several variables began to draw the attention of mathematicians even as the foundations of Mathematical Analysis were taking shape. It spurred on the development of Differential Calculus, and in 1629 Pierre de Fermat obtained the necessary condition for an extremum (i.e., the gradient is zero)—one of the celebrated results in Analysis. He was followed by Isaak Newton and Gottfried Wilhelm von Leibniz, who essentially formulated the second-order conditions for an extremum (i.e., in terms of second derivatives).

Another class of extremum problems that have been traditional among mathematicians, includes problems of variational calculus. They date back to ancient times when isoperimetric problems were examined. However, the real beginning of variational calculus belongs to the end of the eighteenth century when Jean Bernoulli stated his famous brachistochrone problem. In today's language, the classical problem of variational calculus is an infinite-dimensional problem of unconstrained optimization, in which the functional to be minimized has a special (integral) form. Leonhard Euler derived first-order extremum conditions (Euler's equation) and Adrien Marie Legendre and Carl Gustav Jacobi the second-order conditions. It was Karl Weierstrass, in the second half of the nineteenth century, who for the first time posed the crucial question of existence of a solution.

The finite-dimensional as well as the infinite-dimensional problems are good examples of unconstrained minimization problems. Constrained extremum problems have been considered in classical mathematics only for equality constraints. Lagrange's method of multipliers (the eighteenth century) is a first-order necessary extremum condition in both the finite-dimensional and infinite-dimensional problems in the calculus of variations. It is interesting to note that similar conditions for inequality constraints have been obtained only recently. Jean Baptist Fourier, Hermann Minkowski, Hermann Weyl, and other mathematicians studied systems of inequalities proper (not related to minimization problems), and developed a mathematical apparatus which

allows to derive easily extremum conditions in problems with inequality constraints.

The first works on extremum problems with constraints of a general nature appeared in the late 1930s or early 1940s. The origins of those works are diverse. The Chicago group of analysts—Gilbert Bliss, Oskar Bolza, E.J. MacShane, L.N. Graves, M.R. Hestenes, and others—shared interest in finding the most general statement of variational problems. A paper of F. Valentine published in Chicago in 1937, dealt with extremum conditions for problems in the calculus of variations, with inequality constraints of various kinds. Then, Edward James MacShane and ~~David Roxbee~~  
Cox developed general schemes for analyzing abstract extremum problems. A graduate student at the University of Chicago, William Karush, did research on finite-dimensional minimization problems with general constraints. In 1939, he derived first-and second-order conditions for an extremum in the smooth case; however, his results went ignored and the work was not published. During the next decade, the American mathematician Fritz John studied extremum problems in geometry (for example, the problem of finding the smallest ellipsoid circumscribing a given convex body) and obtained essentially the same extremum conditions. But a notable mathematical journal did not accept John's work for publication, and it first appeared only in 1949.

Independently of American researchers, Soviet mathematicians made their contribution to the study of optimization problems. Leonid Vital'evich Kantorovich is a pioneer in this field of mathematics. In 1939 he formulated a number of problems in economics, which were well beyond the standard mathematical apparatus—they were problems of minimization of a linear function on a set given by linear constraints in the form of equalities as well as of inequalities. Kantorovich developed the theory and the methods (not entirely algorithmic) for solving them. In 1940, Kantorovich published an article in which he gave a general formulation of extremum conditions with constraints in an infinite-dimensional space. However, Kantorovich's work did not stir the mathematical community of that time, and remained practically unnoticed. As the reader may observe, fate was not kind to those who pioneered in the study of nonclassical optimization problems.

The situation changed in the late 1940s. During the World War II the American mathematician George B. Dantzig, being involved in industrial applications, studied problems of minimizing a linear function under linear constraints, which became known as "linear programming" problems. Dantzig formulated conditions for optimality of solutions in linear programming. Inspired by John Von Neumann's work in game theory, Dantzig, David Gale and later Harold William Kuhn and Albert William Tucker developed duality theory in linear programming—a specific formulations of extremum conditions.

In the wake of the linear programming theory, its natural generalization to the nonlinear case unfolded. The problem of minimizing a nonlinear function

under nonlinear constraints became known as the mathematical programming problem—hardly a well-chosen term because of the enormous scope subsumed by both adjectives. When the objective function and the constraints are convex the problem is referred to as a convex programming problem. Extremum conditions for mathematical programming problems became widely known after Kuhn and Tucker published their results in 1950. They obtained essentially the same results as William Karush and Fritz John; however, they formulated extremum conditions in terms of a saddle point for the convex case, which is applicable as well in the nonsmooth case.

The so-called optimal control problems were the next step in developing the theory of optimization. These problems are an immediate generalization of the classical problem of the calculus of variations. They consist in optimization of functionals of solutions of usual differential equations, the right-hand sides of which contain functions subject to choice ("controls"). L.S. Pontryagin, V.G. Boltyanskij, and R.V. Gamkrelidze stated and proved necessary optimality conditions for these problems as the so-called maximum principle (1956-58). In a different form, optimality conditions were obtained by Richard E. Bellman, who used the concepts of dynamic programming. His results concerned a very specific form of optimal control problems, and the fact that they were related to extremum conditions for mathematical programming problems was not recognized at that time.

In the 1960s, A.Ya. Dubovitskij and A.A. Milyutin, and also B.N. Pshenichnyj, Lucien W. Neustadt, Hubert Halkin, Jack Warga, among others, delineated general techniques for obtaining extremum conditions for abstract optimization problems with constraints as to include both the Kuhn-Tucker theorem and the maximum principle. This enabled mathematicians to review the current results and, in particular, to divide them into two groups: (1) standard results to be obtained through general techniques and (2) nonstandard results which depend on a particular problem. Convex analysis turned out to be a convenient tool for investigating extremum problems; this recent part of mathematics has been perfected by ~~Richard~~ Rockafellar and other mathematicians. - Terry

So far we have spoken only of extremum conditions in the theory of optimization. However the extremum conditions are inadequate to provide an explicit solution of the problem. Soon it became clear that it was difficult, if not impossible, to find analytic solutions at all, and one has a choice to be satisfied with an algorithmic solution—an iterative algorithm, which, in principle, can approximate the solution to any required degree of accuracy. This was a fundamentally new view. The emergence and development of digital computers further bolstered this approach and led to changes in optimization problematics. Numerical methods for solving optimization problems have become a new area of mathematics: "computing" mathematics.

Computational problems were of little interest to mathematicians of the past centuries. Some methods for solving nonlinear equations and methods of

unconstrained minimization are associated with names such as Isaak Newton, Carl Friedrich Gauss, Augustin-Louis Cauchy, but the results that they and other mathematicians who came after them obtained, remained for long time obscure or sporadic.

Perhaps statisticians were the first who felt the need for numerical minimization methods. In solving parameter estimation problems, the maximum likelihood method, or the least squares method, called for finding an extremum of a function of many variables (in general, nonquadratic function). In the 1940s through the 1950s, statisticians, e.g., Haskell Curry, Kenneth Levenberg, Earl David Crocket, Herman Chernoff, made the first steps in investigating numerical methods of unconstrained minimization. In the early 1950s, David Cox, Herbert Robbins and Sutton Monro, Jack Kiefer and Jacob Wolfowitz developed methods for minimizing functions in random noise, in solving problems of experiment design or regression equations.

Linear algebra was another area of mathematics in which optimization methods took its rise. Solving large systems of linear equations in the case of the finite-difference approximation of partial differential equations entailed the development of iterative methods of linear algebra. However, the problem of solving a system of linear equations is equivalent to that of minimizing a quadratic function, and many methods are convenient to construct and prove on the basis of this fact. These are the method of componentwise descent, the steepest descent method, the conjugate-gradient method, and some other methods of linear algebra. It was only natural to extend these methods to the nonquadratic case.

Specialists in automatic control theory, too, were faced with the need to solve optimization problems. In the 1950s, V.V. Kazakevich, A.A. Feldbaum, and A.A. Pervozvanskij developed the theory of extremum control and special optimization methods for dynamic systems in the real-time.

The first numerical method of nonlinear programming—the penalty-function method—was introduced by Richard Courant in 1943. The method was based on the physical considerations of the problem in question. The simplex method was suggested by Dantzig in the late 1940s to solve linear programming problems, and gave an impetus to a further development of optimization methods. Abundance of applications and efficient computer programs made the simplex method popular, especially with economists.

Initially, research in optimization methodology was sporadic and involved neither a unified nor any definite methodology. However, in the mid-1960s, a definitive trend developed in computational mathematics dealing with numerical optimization methods. New methods were developed, and new classes of problems were examined. At the same time, a unified mathematical apparatus was constructed to analyze the convergence, including the rate of convergence, and optimization methods have been well defined and classified. Today, optimization methodology is quite elaborate, and covers all the basic classes of optimization problems: problems of unconstrained minimization of smooth

and nonsmooth functions in finite-dimensional and infinite-dimensional spaces, problems of constrained minimization with equality and/or inequality constraints in both the convex and nonconvex cases, etc. Rigorous proofs have been constructed for most of the methods, the rate of convergence has been defined, the range of applications has been outlined. Of course, many problems have not been yet completely solved. New, efficient methods are needed for problems in specific applications, accessible and well-tested software has to be developed; this constitutes only part of what still has to be done.

It seems to us that the numerical optimization methods have now matured. The objective of this book is to present in the systematic order the “state of the art” of optimization.

## TRANSLITERATION TABLE (RUSSIAN-ENGLISH)

R	E	R	E
а А	a	р Р	r
б Б	b	с С	s
в В	v	т Т	t
г Г	g	у У	u
д Д	d	ф Ф	f
е Е	e	х Х	kh
ё Ё	e	ц Ц	ts
ж Ж	zh	ч Ч	ch
з З	z	ш Ш	sh
и И	i	щ Щ	shch
й Й	j	ъ Ъ	"
к К	k	ы ы	y
л Л	l	ь ь	'
м М	m	э Э	eh
н Н	n	ю Ю	yu
о О	o	я Я	ya
п П	p		

## NOTATION

$\mathbf{R}^n$  is  $n$ -dimensional real Euclidean space;

$\{x_1, \dots, x_n\}$  are the components of the vector  $x \in \mathbf{R}^n$ ;  $\mathcal{L}_1$

$\|\cdot\|$  is the norm in  $\mathbf{R}^n$ :  $\|x\|^2 = x_1^2 + \dots + x_n^2$ .

$(\cdot, \cdot)$  is the scalar product in  $\mathbf{R}^n$ :  $(x, y) = x_1y_1 + \dots + x_ny_n$ ;

$I$  is the identity matrix;

$A^T$  is the transpose of the matrix  $A$ ;

$A^+$  is the pseudoinverse of the matrix  $A$  (Sec. 6.1);

$A \geq B$ : matrices  $A$  and  $B$  are symmetric and  $A - B$  is nonnegative definite;

$A > B$ : matrices  $A$  and  $B$  are symmetric and  $A - B$  is positive definite;

$\|A\|$  is the norm of the matrix  $A$ :  $\|A\| = \max_{\|x\|=1} \|Ax\|$ ;

$\rho(A)$  is the spectral radius of the matrix  $A$  (Sec. 2.1);

$x \geq y$ : all components of  $x \in \mathbf{R}^n$  are not less than the corresponding components of  $y \in \mathbf{R}^n$ ,  $x_i \geq y_i$ ,  $i = 1, \dots, n$ ;

$\mathbf{R}_+^n$  is the nonnegative orthant in  $\mathbf{R}^n$ :  $\mathbf{R}_+^n = \{x \in \mathbf{R}^n : x \geq 0\}$ ;

$x_+$  is the positive part of  $x \in \mathbf{R}^n$ :  $(x_*)_i = \max \{0, x_i\}$ ,  $i = 1, \dots, n$ ;

$x^* = \underset{x \in Q}{\operatorname{argmin}} f(x)$  is any global minimum point of  $f(x)$  on  $Q$ :  $x^* \in Q$ ,

$$f(x^*) = \underset{x \in Q}{\operatorname{min}} f(x);$$

$X^* = \underset{x \in Q}{\operatorname{argmin}} f(x)$  is the set of global minimum points of  $f(x)$  on  $Q$ :  $\mathcal{A}$

$$X^* = \{x^* = \underset{x \in Q}{\operatorname{argmin}} f(x)\}; \quad \text{span style="float: right;"> $\mathcal{S}$$$

$\nabla f(x)$ ,  $f'(x)$  is the gradient of the scalar function  $f(x)$  (Sec. 1.1);

$\nabla g(x)$ ,  $g'(x)$  is the derivative of the vector function  $g(x)$ , the Jacobi matrix (Sec. 1.1);

$\nabla^2 f(x)$ ,  $f''(x)$  is the matrix of second derivatives, the Hessian (Sec. 1.1);

- $\vdash''$   $L'_x(x,y)$ ,  $L''_{xx}(x,y)$ : the gradient and matrix of second derivatives of  $L(x,y)$  with respect to  $x$ ;  
 $\partial f(x)$ : the subgradient of the convex function (Secs. 5.1 and 9.1);  
 $\partial_\varepsilon f(x)$ : the  $\varepsilon$ -subgradient of the convex function (Sec. 5.1);  
 $f'(x|y)$ : the derivative of  $f(x)$  at the point  $x$  in the direction  $y$  (Secs. 1.1 and 5.1);  
 $D(f)$  is the domain of definition of  $f(x)$  (Sec. 5.1);  
 $\text{Cov } Q$  is the convex hull of the set  $Q$  (Sec. 5.1);  
 $Q^0$  is the interior of  $Q$ ;  
 $\emptyset$  is the empty set;  
 $P_Q(x)$  is the projection of the point  $x$  onto the set  $Q$  (Sec. 5.1);  
 $\rho(x,Q)$  is the distance from the point  $x$  to the set  $Q$ :  $\rho(x,Q) = \inf_{y \in Q} \|x - y\|$   
 $o(h(x))$ : if  $g: \mathbf{R}^n \rightarrow \mathbf{R}^m$ ,  $h: \mathbf{R}^n \rightarrow \mathbf{R}^s$  and  $\|g(x)\|/\|h(x)\| \rightarrow 0$  as  $\|x\| \rightarrow 0$ , then  $g(x) = o(h(x))$ ;  
 $O(h(x))$ : if  $g: \mathbf{R}^n \rightarrow \mathbf{R}^m$ ,  $h: \mathbf{R}^n \rightarrow \mathbf{R}^s$  and there are  $\varepsilon > 0$ ,  $\alpha$  such that  $\|g(x)\| \leq \alpha \|h(x)\|$  for  $\|x\| \leq \varepsilon$ , then  $g(x) = O(h(x))$ ;  
 $o(u_k)$ : if the sequences  $u_k \in \mathbf{R}^n$ ,  $v_k \in \mathbf{R}^m$ ,  $k = 1, 2, \dots$ , are such that  $\|v_k\|/\|u_k\| \rightarrow 0$  as  $k \rightarrow \infty$ , then  $v_k = o(u_k)$ ;  
 $\mathcal{O}_{k_0} O(u_k)$ : if for sequences  $u_k \in \mathbf{R}^n$ ,  $v_k \in \mathbf{R}^m$ ,  $k = 1, 2, \dots$ , there are  $\alpha > 0$ ,  $k_0$  such that  $\|v_k\| \leq \alpha \|u_k\|$  for  $k \geq k_0$ , then  $v_k = O(u_k)$ ;  
 $E\xi$  is the mathematical expectation of the random variable  $\xi$ ;  
 $E(\xi|x)$  is the conditional mathematical expectation of the random variable  $\xi$  depending on  $x$  for a fixed value of  $x$ ;  
 $\forall \bar{x}$   $\forall$  is the universal quantifier:  $\forall x \in Q$  means “for all  $\bar{x} \in Q$ ";  
 $\square$  is the sign put at the end of a proof (or at the end of an assertion if it is given without proof).  
 Usually the letters  $x$ ,  $y$ ,  $a$ ,  $b$  are used for vectors;  $\alpha$ ,  $\beta$ , ... for scalars;  $A$ ,  $B$ , ... for matrices;  $i$ ,  $j$ ,  $k$ , ... for integers;  $Q$ ,  $S$ , ... for sets. An iterative sequence of vectors is written  $x^0, x^1, \dots, x^k, \dots$ ;  $x_i$  are the components of the vector  $x$ .

## **PART I**

### **UNCONSTRAINED MINIMIZATION**

## CHAPTER 1

### FUNDAMENTALS OF THE THEORY AND METHODS OF UNCONSTRAINED MINIMIZATION

We begin our study of optimization problems with the classical problem of unconstrained minimization of a smooth function:  $\min f(x)$ ,  $x \in \mathbf{R}^n$ .

We focus our attention on this problem not only because of its importance, but also because, due to its simplicity, it clearly exhibits the main features of the nature of optimization problems and theoretical foundations thereof.

## 1.1 REVIEW OF MATHEMATICAL ANALYSIS

### 1.1.1 Differentiation of Scalar Functions

A scalar function  $f(x)$  of an  $n$ -dimensional argument  $x$  ( $f: \mathbf{R}^n \rightarrow \mathbf{R}^1$ ) is said to be *differentiable at a point  $x$*  if we can find a vector  $a \in \mathbf{R}^n$  such that for all  $y \in \mathbf{R}^n$ ,

$$f(x + y) = f(x) + (a, y) + o(y). \quad (1)$$

The vector  $a$  in (1) is called the *derivative* or the *gradient* of  $f(x)$  at a point  $x$  and is written  $f'(x)$  or  $\nabla f(x)$ . Thus, the gradient is defined by

$$f(x + y) = f(x) + (\nabla f(x), y) + o(y). \quad (2)$$

In other words, a function is differentiable at a point  $x$  if it admits a first-order linear approximation at  $x$ , i.e., we can find a linear function  $\tilde{f}(y) = f(x) + (\nabla f(x), y)$  such that  $f(x + y) - \tilde{f}(y) = o(y)$ . It is clear that the gradient is uniquely determined,  $\nabla f(x)$  being a vector with components  $(\partial f(x)/\partial x_1, \dots, \partial f(x)/\partial x_n)$ . One can calculate the gradient directly

from the definition; or by using its coordinate form; or by using the rule of differentiating a composite function (see (12)).

For example, let  $f(x)$  be the quadratic function

$$f(x) = (Ax, x)/2 - (b, x),$$

where  $A$  is a symmetric  $n \times n$ -matrix,  $b \in \mathbf{R}^n$ . Then

$$\begin{aligned} f(x+y) &= (A(x+y), x+y)/2 - (b, (x+y)) \\ &= (Ax, x)/2 - (b, x) + (Ax-b, y) + (Ay, y)/2 \\ &= f(x) + (Ax-b, y) + (Ay, y)/2. \end{aligned}$$

But  $|(Ay, y)| \leq \|A\| \|y\|^2$ . Hence  $(Ay, y)/2 = o(y)$ . Thus,  $f(x)$  is differentiable at any point  $x$  and

$$\nabla f(x) = Ax - b. \quad (3)$$

The function  $f(x)$  is said to be *differentiable on a set  $Q \subset \mathbf{R}^n$*  if it is differentiable at all points of  $Q$ . If  $f(x)$  is differentiable on the entire space  $\mathbf{R}^n$ , then it is said to be simply *differentiable*.

Suppose  $f(x)$  is differentiable on the segment  $[x, x+y]$  (i.e., for points of the form  $x + \tau y$ ,  $0 \leq \tau \leq 1$ ). We consider the one-variable function  $\phi(\tau) = f(x + \tau y)$  and compute its derivative for  $0 \leq \tau \leq 1$ :

$$\frac{\phi(\tau + \Delta\tau) - \phi(\tau)}{\Delta\tau} = \frac{f(x + (\tau + \Delta\tau)y) - f(x + \tau y)}{\Delta\tau}$$

$$= \frac{(\nabla f(x + \tau y), \Delta\tau y) + o(\Delta\tau y)}{\Delta\tau},$$

$$\phi'(\tau) = \lim_{\Delta\tau \rightarrow 0} \frac{\phi(\tau + \Delta\tau) - \phi(\tau)}{\Delta\tau} = (\nabla f(x + \tau y), y).$$

Thus,  $\phi(\tau)$  is differentiable on  $[0,1]$  and

$$\phi'(\tau) = (\nabla f(x + \tau y), y). \quad (4)$$

The quantity

$$f'(x; y) = \lim_{\varepsilon \rightarrow +0} \frac{f(x + \varepsilon y) - f(x)}{\varepsilon} \quad (5)$$

is called the *directional derivative* (or *variation*) of  $f(x)$  at  $x$  in the direction  $y$ . The directional derivative may exist for nonsmooth functions as well. For example, for  $f(x) = \|x\|$  we have  $f'(0; y) = \|y\|$ . If  $f(x)$  has a derivative linear in  $y$  in all directions at a point  $x$ :  $f'(x; y) = (a, y)$ , then  $f(x)$  is *Gâteaux differentiable* at the point  $x$ . Such a function has partial derivatives,  $f'(x; e_i) = \partial f(x)/\partial x_i$  ( $e_i$  are the coordinate basis vectors),  $a = (\partial f/\partial x_1, \dots, \partial f/\partial x_n)$ . It follows from formula (4) that if  $f(x)$  is differentiable at  $x$ , then it is also Gâteaux differentiable, with

$$f'(x; y) = \phi'(0) = (\nabla f(x), y). \quad (6)$$

The converse does not generally hold. For example, the function  $f: \mathbf{R}^n \rightarrow \mathbf{R}^1$ ,  $n \geq 2$ , of the form

$$f(x) = \begin{cases} 1 & \text{if } \|x - a\| = \|a\|, x \neq 0, \\ 0 & \text{otherwise,} \end{cases} \quad (7)$$

where  $a \in \mathbf{R}^n$ ,  $a \neq 0$ , is differentiable at zero in any direction and  $f'(0; y) = 0$  for all  $y$ , i.e., it is Gâteaux differentiable at zero, yet is not differentiable (and not even continuous) at zero. Sometimes, to emphasize the difference from the Gâteaux differentiability, the term *Fréchet differentiability* rather than *differentiability* is used.

If a function  $f(x)$  is differentiable on  $[x, x+y]$ , then, using (4) and the Newton-Leibniz formula

$$\phi(1) = \phi(0) + \int_0^1 \phi'(\tau) d\tau,$$

we obtain an expression for the remainder in (2) in the integral form:

$$\begin{aligned} f(x+y) &= f(x) + \int_0^1 (\nabla f(x+\tau y), y) d\tau \\ &= f(x) + (\nabla f(x), y) + \int_0^1 (\nabla f(x+\tau y) - \nabla f(x), y) d\tau. \end{aligned} \quad (8)$$

Another useful result—the mean value theorem—follows from the finite-increment formula  $\phi(1) = \phi(0) + \phi'(\theta)$ ,  $0 \leq \theta \leq 1$ , and from (4):

$$f(x+y) = f(x) + (\nabla f(x+\theta y), y), \quad (9)$$

where  $0 \leq \theta \leq 1$  is some number.

## Exercises

1. Prove:

- (a)  $\nabla \|x\| = x/\|x\|$  for  $x \neq 0$ ; for  $x = 0$  the function  $\|x\|$  is nondifferentiable;
  - (b)  $\nabla \|x_+\|^2 = 2x_+$ .
2. Prove that continuity in  $x$  of the Gâteaux derivative implies differentiability.

### 1.1.2 Differentiation of Vector Functions

We have considered so far the differentiability of scalar functions. The vector-function version is defined analogously. The function  $g: \mathbf{R}^n \rightarrow \mathbf{R}^m$  is said to be *differentiable at a point  $x$*  if we can find an  $m \times n$ -matrix  $A$  such that for all  $y \in \mathbf{R}^n$ ,

$$g(x + y) = g(x) + Ay + o(y). \quad (10)$$

The matrix  $A$  is called the *derivative* or the *Jacobian matrix* of the mapping  $g(x)$  and is denoted the same as in the scalar case,  $g'(x)$  or  $\nabla g(x)$ . Thus

$$g(x + y) = g(x) + g'(x)y + o(y), \quad (11)$$

i.e., a function differentiable at  $x$  admits at  $x$  a first-order linear approximation. Obviously, for a differentiable vector function  $g(x) = (g_1(x), \dots, g_m(x))$  the elements of the Jacobian matrix are defined by the formula  $g'(x)_{ij} = \partial g_i(x)/\partial x_j$ .

If  $m = 1$ , then  $g'(x)$  is a  $1 \times n$ -matrix, i.e., a row vector. It is more convenient, however, to assume all vectors to be column ones; taking this into account, definition (2) was adopted, where  $\nabla f(x)$  is a column vector. There is no ambiguity, but one needs to be careful when applying general formulas to the case  $m = 1$ , and should use transposition if necessary.

Let  $g: \mathbf{R}^n \rightarrow \mathbf{R}^m$  be differentiable at  $x$ , and  $h: \mathbf{R}^m \rightarrow \mathbf{R}^s$  be differentiable at  $g(x)$ . Then the *rule for differentiation of composite functions* (or *chain rule*) is valid:

$$[h(g(x))]' = h'(g(x)) g'(x), \quad (12)$$

where the right-hand side contains the product of the matrices  $h'$  and  $g'$ .

The mean value theorem does not hold for vector functions, i.e., there does not generally exist  $\theta$ ,  $0 \leq \theta \leq 1$ , such that

$$g(x + y) = g(x) + g'(x + \theta y)y$$

for a function  $g: \mathbf{R}^n \rightarrow \mathbf{R}^m$ ,  $m > 1$ , differentiable on  $[x, x+y]$ . But there is a formula analogous to (8): if  $g(x)$  is differentiable on  $[x, x+y]$ , then

$$\begin{aligned} g(x+y) &= g(x) + \int_0^1 g'(x+\tau y)y \, d\tau \\ &= g(x) + g'(x)y + \int_0^1 (g'(x+\tau y) - g'(x))y \, d\tau \end{aligned} \quad (13)$$

yielding, in particular, the following useful estimates. If  $\|g'(x+\tau y)\| \leq L$  for  $0 \leq \tau \leq 1$ , then

$$\|g(x+y) - g(x)\| \leq L\|y\|, \quad (14)$$

whereas, if  $g'(x)$  satisfies a Lipschitz condition on  $[x, x+y]$ :

$$\|g'(u) - g'(v)\| \leq L\|u - v\|, \quad u, v \in [x, x+y],$$

then

$$\|g(x+y) - g(x) - g'(x)y\| \leq L\|y\|^2/2. \quad (15)$$

As in the scalar case, a function  $g: \mathbf{R}^n \rightarrow \mathbf{R}^m$  differentiable at all points of  $\mathbf{R}^n$  is called *differentiable*.

### Exercise

3. Using (12) and the result of Problem 1, prove that

$$\nabla\|(Ax - b)_+\|^2 = 2A^T(Ax - b)_+,$$

where  $A$  is an  $m \times n$ -matrix.

#### 1.1.3 Second Derivatives

A scalar function  $f(x)$  on  $\mathbf{R}^n$  is said to be *twice differentiable at a point  $x$*  if it is differentiable at  $x$  and we can find a symmetric  $n \times n$ -matrix  $H$  such that for all  $y \in \mathbf{R}^n$ ,

$$\sqrt{\|y\|^2} \quad f(x+y) = f(x) + (\nabla f(x), y) + (Hy, y)/2\sqrt{\|y\|^2}$$

This matrix is called *the matrix of second derivatives*, the *Hessian matrix*, or the *Hessian*, and is denoted  $f''(x)$  or  $\nabla^2 f(x)$ . In other words, a function is twice differentiable at a point  $x$  if it admits a second-order quadratic ap-

proximation in a neighborhood of the point  $x$ , i.e., there exists a quadratic function

$$\tilde{f}(y) = f(x) + (\nabla f(x), y) + (\nabla^2 f(x)y, y)/2$$

such that

$$|f(x+y) - \tilde{f}(y)| = o(\|y\|^2).$$

Let us sharpen the estimates obtained earlier for twice-differentiable functions. We again consider the scalar function  $\phi(\tau) = f(x + \tau y)$ , assuming that  $f$  is twice differentiable on  $[x, x+y]$ . As above, we show that this function is twice differentiable and

$$\phi''(\tau) = (\nabla^2 f(x + \tau y)y, y). \quad (17)$$

Then from the Taylor formula with the integral remainder

$$\phi(1) = \phi(0) + \phi'(0) + \int_0^1 \int_0^t \phi''(\tau) d\tau dt$$

we obtain

$$f(x+y) = f(x) + (\nabla f(x), y) + \int_0^1 \int_0^t (\nabla^2 f(x+\tau y)y, y) d\tau dt. \quad (18)$$

In particular, if

$$\|\nabla^2 f(x + \tau y)\| \leq L, \quad 0 \leq \tau \leq 1,$$

we have

$$|f(x+y) - f(x) - (\nabla f(x), y)| \leq (L/2)\|y\|^2, \quad (19)$$

whereas if

$$\|\nabla^2 f(x + \tau y) - \nabla^2 f(x)\| \leq L\tau\|y\|,$$

then

$$|f(x+y) - f(x) - (\nabla f(x), y) - (\frac{1}{2})(\nabla^2 f(x)y, y)| \leq (L/6)\|y\|^3. \quad (20)$$

If we use the Taylor formula with remainder in the Lagrange form,

$$\phi(1) = \phi(0) + \phi'(0) + \phi''(\theta)/2, \quad 0 \leq \theta \leq 1,$$

then we can find a  $\theta$ ,  $0 \leq \theta \leq 1$ , such that

$$f(x+y) = f(x) + (\nabla f(x), y) + (\nabla^2 f(x) + \theta y)y/2.$$



## Exercises

4. Show that  $\nabla^2 f(x)$  is the matrix with elements  $\partial^2 f(x)/\partial x_i \partial x_j$ .

5. Prove:

(a)  $\nabla^2[(Ax, x)/2 - (b, x)] \equiv A$ , where  $A$  is a symmetric  $n \times n$ -matrix,  $b \in \mathbf{R}^n$ ;

$$(b) \nabla^2 \|x\| = I \|x\|^{-1} - xx^T \|x\|^{-3}, x \neq 0;$$

$$(c) \nabla^2(c, x)^2 = 2cc^T, c \in \mathbf{R}^n.$$

✓' 6. Check that  $f''(x) = (f'(x))'$ , i.e., the derivative of the vector function  $f'(x)$  coincides with the second derivative of  $f(x)$ .

### 1.1.4 Convex Functions

The notion of convexity plays a significant role in extremum theory, and we will often employ it. A scalar function  $f(x)$  on  $\mathbf{R}^n$  is said to be *convex* if

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y) \quad (22)$$

for any  $x, y \in \mathbf{R}^n$ ,  $0 \leq \lambda \leq 1$ . This definition has an intuitive geometric interpretation: the graph of the function on the segment  $[x, y]$  lies below the chord joining the points  $(x, f(x))$  and  $(y, f(y))$  (Fig. 1). The definition of convexity involves pairs of points  $x, y$  and their convex combinations. A similar inequality holds for convex combinations of any number of points.

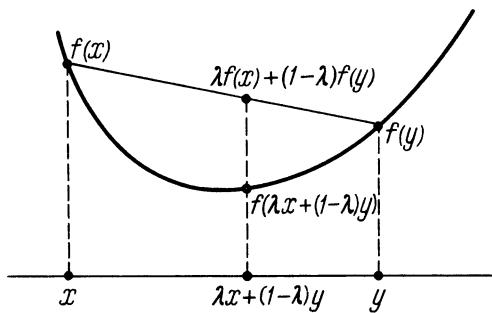


Fig. 1 A convex function.

**LEMMA 1** (Jensen's inequality). Let  $f(x)$  be a convex function on  $\mathbf{R}^n$ . Then for any  $x^1, \dots, x^k \in \mathbf{R}^n$  and  $\lambda_i \geq 0$ ,  $i = 1, \dots, k$ ,  $\sum_{i=1}^k \lambda_i = 1$ , one has

$$f(\lambda_1 x^1 + \dots + \lambda_k x^k) \leq \lambda_1 f(x^1) + \dots + \lambda_k f(x^k). \quad \square \quad (23)$$

A function  $f(x)$  such that  $-f(x)$  is convex is called *concave*. Obviously, the *affine* function  $f(x) = (a, x) + \beta$  is both convex and concave.

It is obvious from the definition that if the  $f_i(x)$  are convex,  $i = 1, \dots, m$ , then  $f(x) = \sum_{i=1}^m \gamma_i f_i(x)$ ,  $\gamma_i \geq 0$ , and  $f(x) = \max_{1 \leq i \leq m} f_i(x)$  are also convex.

Strictly and strongly convex functions are an important special case of convex functions. A function  $f(x)$  on  $\mathbf{R}^n$  is called *strictly convex* if for any  $x \neq y$ ,  $0 < \lambda < 1$ ,

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y), \quad (24)$$

and is called *strongly convex with constant  $\ell > 0$*  if for  $0 \leq \lambda \leq 1$ ,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \ell\lambda(1-\lambda)\|x - y\|^2/2. \quad (25)$$

Clearly, a strongly convex function is strictly convex.

It is important to have analytic criteria to evaluate whether a function is convex or not. Such criteria exist and are simplest for differentiable functions. They are based on the following elementary result.

**LEMMA 2.** Let  $\psi(t)$  be a differentiable function on  $\mathbf{R}^1$ . Then the convexity of  $\psi(t)$  is equivalent to the monotonicity of the derivative ( $\psi'(\tau_1) \geq \psi'(\tau_2)$  for  $\tau_1 \geq \tau_2$ ), strict convexity to strict monotonicity ( $\psi'(\tau_1) > \psi'(\tau_2)$  for  $\tau_1 > \tau_2$ ), and the strong convexity to the strong monotonicity of ( $\psi'(\tau_1) - \psi'(\tau_2) \geq \ell(\tau_1 - \tau_2)$ ,  $\tau_1 > \tau_2$ ).  $\square$

**LEMMA 3.** For a differentiable function  $f(x)$  on  $\mathbf{R}^n$ , convexity is equivalent to the inequality

$$f(x + y) \geq f(x) + (\nabla f(x), y), \quad (26)$$

strict convexity to the inequality

$$f(x + y) > f(x) + (\nabla f(x), y), \quad y \neq 0, \quad (27)$$

and strong convexity to the inequality

$$f(x + y) \geq f(x) + (\nabla f(x), y) + \ell\|y\|^2/2 \quad (28)$$

for any  $x, y \in \mathbf{R}^n$ .  $\square$

In other words, the graph of a (strictly) convex function lies (strictly) above the tangent hyperplane, whereas for a strongly convex function the graph lies above some paraboloid (Fig. 2).

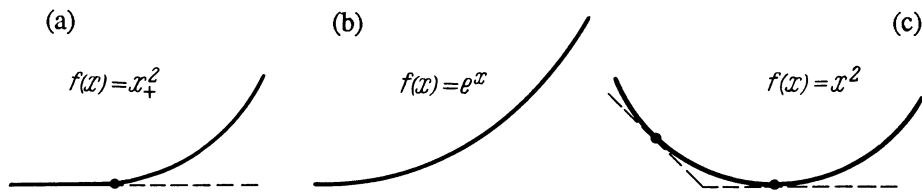


Fig. 2 Types of convexity: (a) a convex function;  
 (b) a strictly convex function;  
 (c) a strongly convex function.

From (26) we obtain the useful inequality

$$(\nabla f(x) - \nabla f(y), x-y) \geq 0 , \quad (29)$$

which is a generalization of the monotonicity condition for the derivative of a convex function to the multidimensional case. For a strictly convex function the *strict monotonicity condition* holds:

$$(\nabla f(x) - \nabla f(y), x-y) > 0 , \quad x \neq y ; \quad (30)$$

for a strongly convex function the *strong monotonicity condition* holds:

$$(\nabla f(x) - \nabla f(y), x-y) \geq \ell \|x-y\|^2 . \quad (31)$$

A criterion for convexity is simplest for twice-differentiable functions \$f(x)\$: convexity is equivalent to the condition

$$\nabla^2 f(x) \geq 0 , \quad (32)$$

and strong convexity is equivalent to the condition

$$\nabla^2 f(x) \geq \ell I \quad (33)$$

for all \$x\$. If

$$\nabla^2 f(x) > 0 \quad (34)$$

for all \$x\$, then \$f(x)\$ is strictly convex. The last condition is only sufficient (for example, for a strictly convex function \$f(x) = \|x\|^4\$ one has \$\nabla^2 f(0) = 0\$).

Let \$x^\*\$ be a minimum point of a differentiable strongly convex function \$f(x)\$ (with constant \$\ell\$). Such a point exists, is unique and \$\nabla f(x^\*) = 0\$ (see Sections 1.2 and 1.3 below). Hence, from inequalities (28), (31) we have

$$f(x) \geq f(x^*) + \ell \|x - x^*\|^2/2 . \quad (35)$$

$$(\nabla f(x), x - x^*) \geq \ell \|x - x^*\|^2 , \quad (36)$$

$$\|\nabla f(x)\| \geq \ell \|x - x^*\| . \quad (37)$$

## Exercise

7. Prove:

- (a) the function  $(Ax, x)/2 - (b, x)$ ,  $A > 0$ , is strongly convex;
- (b) the function  $(Ax, x)/2 - (b, x)$  with singular matrix  $A \geq 0$  (in particular, a linear function) is convex, but not strictly convex;
- (c) the function  $\|x\|^\alpha$  is convex for  $\alpha \geq 1$ , strictly convex for  $\alpha > 1$ , strongly convex only for  $\alpha = 2$ .

## 1.2 EXTREMUM CONDITIONS

Extremum conditions for smooth functions on the entire space are well known. We will, however, consider them in some detail, since they can be used as a model for constructing similar conditions in more complex cases.

### 1.2.1 A First-order Necessary Condition

The point  $x^*$  is called a *local minimum* of  $f(x)$  on  $\mathbf{R}^n$  if we can find an  $\varepsilon > 0$  such that  $f(x) \geq f(x^*)$  for all  $x$  in an  $\varepsilon$ -neighborhood of  $x^*$  (i.e., for  $\|x - x^*\| \leq \varepsilon$ ). In this case, one sometimes calls  $x^*$  simply a *minimum point*. However, one needs to bear in mind the distinction between a local minimum point and a *global* minimum point (i.e., a point  $x^*$  such that  $f(x) \geq f(x^*)$  for all  $x$ ). In necessary conditions for an extremum, one can simply speak of a minimum point, since some property holds for a local minimum as well as for a global minimum. In formulating sufficient conditions, the distinction has to be made as to which kind of a minimum point is involved.

**THEOREM 1** (Fermat). Let  $x^*$  be a minimum point of  $f(x)$  on  $\mathbf{R}^n$  and let  $f(x)$  be differentiable at  $x^*$ . Then

$$\nabla f(x^*) = 0 . \quad (1)$$

**PROOF.** Suppose  $\nabla f(x^*) \neq 0$ . Then

$$\begin{aligned} f(x^* - \tau \nabla f(x^*)) &= f(x^*) - \tau \|\nabla f(x^*)\|^2 + o(\tau \nabla f(x^*)) \\ &= f(x^*) - \tau (\|\nabla f(x^*)\|^2 + \tau^{-1} o(\tau)) < f(x^*) \end{aligned}$$

for sufficiently small  $\tau > 0$  by the definition of  $o(\tau)$ . But this contradicts the fact that  $x^*$  is a local minimum point.  $\square$

This proof is very instructive. Under the assumption that the extremum condition is not satisfied, we showed how to construct a point with a smaller value of  $f(x)$ . Thus, this proof illustrates the way to construct a minimization method. This method (known as the gradient method) will be examined in detail in Section 1.4.

### 1.2.2 A First-order Sufficient Condition

Certainly, even if some point happens to be *stationary* (i.e., the gradient vanishes at this point), it need not be a minimum point (Fig. 3). For example, it can be a maximum point or a saddle point. For convex functions, though, this situation is impossible.

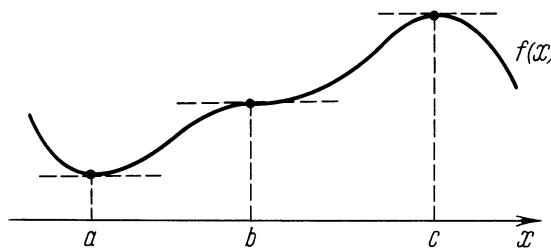


Fig. 3 Stationary points:  $a$  is a minimum point,  
 $b$  is an inflection point;  $c$  is a maximum point.

**THEOREM 2.** Let  $f(x)$  be a convex function differentiable at a point  $x^*$  and let  $\nabla f(x^*) = 0$ . Then  $x^*$  is a global minimum point of  $f(x)$  on  $\mathbf{R}^n$ .

**PROOF.** The proof follows immediately from formula (26) of Section 1.1 since  $f(x) \geq f(x^*) + (\nabla f(x^*), x - x^*) = f(x^*)$  for any  $x \in \mathbf{R}^n$ .  $\square$

Thus, for convex functions the necessary extremum condition is also a sufficient one. Later on we will see that this situation is also common to other types of convex extremum problems.

### 1.2.3 A Second-order Necessary Condition

For nonconvex problems, one can continue the investigation of extremum conditions, using higher derivatives.

**THEOREM 3.** Let  $x^*$  be a minimum point of  $f(x)$  on  $\mathbf{R}^n$  and let  $f(x)$  be twice differentiable at  $x^*$ . Then

$$\nabla^2 f(x^*) \geq 0 . \quad (2)$$

**PROOF.** By Theorem 1,  $\nabla f(x^*) = 0$  and hence for an arbitrary  $y$  and a sufficiently small  $\tau$

$$f(x^*) \leq f(x^* + \tau y) = f(x^*) + \tau^2 (\nabla^2 f(x^*) y, y)/2 + o(\tau^2) ,$$

$$(\nabla^2 f(x^*) y, y) \geq o(\tau^2)/\tau^2 .$$

Passing to the limit as  $\tau \rightarrow 0$ , we obtain  $(\nabla^2 f(x^*) y, y) \geq 0$ . Since  $y$  is arbitrary,  $\nabla^2 f(x^*) \geq 0$ .  $\square$

#### 1.2.4 A Second-order Sufficient Condition

**THEOREM 4.** At a point  $x^*$ , let  $f^*(x)$  be twice differentiable, let a first-order necessary condition hold (i.e.,  $\nabla f(x^*) = 0$ ) and let

$$\nabla^2 f(x^*) > 0 . \quad (3)$$

Then  $x^*$  is a local minimum point.

**PROOF.** Let  $y$  be any vector with unit norm. Then

$$\begin{aligned} f(x^* + \tau y) &= f(x^*) + \tau^2 (\nabla^2 f(x^*) y, y)/2 + o(\tau^2 \|y\|^2) \\ &\geq f(x^*) + \tau^2 \ell/2 + o(\tau^2) , \end{aligned}$$

where  $\ell > 0$  is the smallest eigenvalue of  $\nabla^2 f(x^*)$  and the function  $o(\tau^2)$  does not depend on  $y$ . Hence we can find a  $\tau_0$  such that for  $0 \leq \tau \leq \tau_0$  we have  $\tau^2 \ell/2 \geq o(\tau^2)$ , i.e.,  $f(x^* + \tau y) \geq f(x^*)$ .  $\square$

If the first- and second-order necessary conditions hold at  $x^*$  (i.e.,  $\nabla f(x^*) = 0$ ,  $\nabla^2 f(x^*) \geq 0$ ), but the second-order sufficient condition does not hold (the matrix  $\nabla^2 f(x^*)$  is not positive definite), then  $x^*$  need not be a minimum point (e.g.,  $f(x) = x^3$ ,  $x \in \mathbf{R}^1$ ) and, theoretically, the analysis can be continued using higher derivatives. For the one-dimensional case, the procedure is well known (it is necessary to find the first nonzero derivative); for the multidimensional case computations are more complicated.

### 1.2.5 What Are Extremum Conditions Good For?

In textbooks on Mathematical Analysis the following procedure is usually recommended for seeking extremum points. First find all points satisfying the first-order condition and next check the second-order conditions, choosing minimum points only. Thus, the extremum conditions would appear to be an adequate tool for solving optimization problems.

We emphasize the fact that this is simply not true. Finding a minimum in explicit form by means of extremum conditions is possible only in rare cases—for specially constructed examples (they are usually the ones given in textbooks). The point is that solving the system of equations  $\nabla f(x) = 0$  is no simpler than solving the original problem, and finding an explicit solution is, as a rule, impossible.

Why then are extremum conditions considered and what is the point of giving them so much attention in extremum theory? To be sure, this is partly a vestige of tradition when an analytic representation was viewed as the solution to the problem. More importantly, in our view, extremum conditions provide the basis on which to construct methods of solving optimization problems, and hence their importance. As we will see below, they, first, can yield much useful information about the properties of the extremum, even when we cannot obtain an explicit solution. Secondly, the proof of extremum conditions or the nature of these conditions can show the way to construct optimization methods. We have seen above that the proof of the condition  $\nabla f(x) = 0$  leads naturally to the gradient method of minimization. Thirdly, in proving the methods, several assumptions have to be made. Also, it is usually required that sufficient condition for an extremum hold at the point  $x^*$ . Thus, extremum conditions appear in theorems on the convergence of methods. Finally, the proofs of convergence are most often based on the fact that the “discrepancy” in the extremum conditions is shown to tend to zero.

## 1.3 EXISTENCE, UNIQUENESS, AND STABILITY OF A MINIMUM

Problems of existence, uniqueness, and stability of a solution are an important part of mathematical theory of extremum problems (and, in particular, problems of unconstrained optimization).

### 1.3.1 Existence of a Minimum

The question of the existence of a minimum point is usually solved quite simply by means of the following theorem.

**THEOREM 1** (Weierstrass). Let  $f(x)$  be continuous on  $\mathbf{R}^n$  and let the set  $Q_\alpha = \{x: f(x) \leq \alpha\}$  for some  $\alpha$  be nonempty and bounded. Then there exists a global minimum point of  $f(x)$  on  $\mathbf{R}^n$ .

**PROOF.** Let

$$f(x^k) \rightarrow \inf_{x \in \mathbf{R}^n} f(x) < \alpha .$$

Then  $x^k \in Q_\alpha$  for sufficiently large  $k$ . The set  $Q_\alpha$  is closed (by the continuity of  $f(x)$ ) and bounded, i.e. compact; hence the sequence  $x^k$  has a limit point  $x^* \in Q_\alpha$ . It follows from the continuity of  $f(x)$  that

$$f(x^*) = \inf_{x \in \mathbf{R}^n} f(x) ,$$

i.e.,

$$x^* = \arg \min_{x \in \mathbf{R}^n} f(x) . \quad \square$$

The assumption of the boundedness of  $Q_\alpha$  is essential (for example, the functions  $x$  and  $1/(1+x^2)$  are continuous on  $\mathbf{R}^1$  but have no minimum point). In some cases one can prove the existence of a solution in situations not covered by Theorem 1 (see Exercise 2 below).

## Exercises

1. Prove that a differentiable strongly convex function on  $\mathbf{R}^n$  attains its minimum (use inequality (28) of Section 1.1 and Theorem 1).
2. Let  $f(x) = (Ax, x) - (b, x)$ ,  $A \geq 0$ , and let  $f(x)$  be bounded below (e.g.,  $f(x) \geq 0$ ). Prove that  $f(x)$  attains its minimum on  $\mathbf{R}^n$ , although the conditions of Theorem 1 do not generally hold (the set  $Q_\alpha$  is not necessarily bounded).

### 1.3.2 Uniqueness of a Solution

We say that a minimum point is *locally unique* if in some neighborhood of it there are no other minimum points. We say that  $x^*$  is a *nonsingular minimum point* if at the  $x^*$  the sufficient second-order extremum condition holds, i.e.,  $\nabla f(x^*) = 0$ ,  $\nabla^2 f(x^*) > 0$ .

**THEOREM 2.** A nonsingular minimum point is locally unique.

**PROOF.** According to Exercise 6 of Section 1.1,

$$\nabla f(x) = \nabla f(x^*) + \nabla^2 f(x^*)(x-x^*) + o(x-x^*) .$$

Hence

$$\begin{aligned}\|\nabla f(x)\| &= \|\nabla^2 f(x^*)(x-x^*)\| + o(\|x-x^*\|) \\ &\geq \ell \|x-x^*\| + o(\|x-x^*\|) > 0\end{aligned}$$

for sufficiently small  $\|x-x^*\|$ , since for  $\nabla^2 f(x^*) = A > 0$  we have  $\|Ax\| \geq \ell \|x\|$  for all  $x$ , where  $\ell > 0$  is the smallest eigenvalue of  $A$ . Thus, in some neighborhood of  $x^*$  there are no stationary points of  $f(x)$  and therefore no minimum points.  $\square$

For convex functions, the answer to the question of uniqueness of a minimum is easy to obtain.

**THEOREM 3.** A minimum point of a strictly convex function is (globally) unique.

**PROOF.** The proof follows immediately from the definition of strict convexity.  $\square$

### 1.3.3 Stability of a Solution

In practical solution of optimization problems, one is continually faced with the following question. Suppose we have discovered a method for constructing a minimizing sequence. Does it converge to the solution? If, instead of the initial minimization problem, can one assert that the solutions are close? Questions like these are the province of extremum theory and involve the notions of stability and correctness. We will use the term “stability” for optimization problems and leave the term “correctness” for problems not involving optimization (solution of algebraic, integral, operator equations, and the like).

The local minimum point  $x^*$  of  $f(x)$  is called *locally stable* if every *local minimizing sequence* converges to it, i.e., there is a  $\delta > 0$  such that  $f(x^k) \rightarrow f(x^*)$ ,  $\|x^k - x^*\| \leq \delta$  imply  $x^k \rightarrow x^*$ .

**THEOREM 4.** A local minimum point of a continuous function  $f(x)$  is locally stable if it is locally unique.

**PROOF.** Let  $x^*$  be locally unique. Take an arbitrary local minimizing sequence  $x^k$ ,  $\|x^k - x^*\| \leq \delta$ ,  $f(x^k) \rightarrow f(x^*)$ . By the compactness of a unit sphere in  $\mathbf{R}^n$ , one can take a convergent subsequence  $x^{k_i} \rightarrow \bar{x}$ ,  $\|\bar{x} - x^*\| \leq \delta$ . It follows from the continuity of  $f(x)$  that  $f(\bar{x}) = \lim f(x^{k_i}) = f(x^*)$ . Then, however,  $\bar{x} = x^*$  since  $x^*$  is a locally unique minimum point. Since the same is true for any other sequence, the entire sequence  $x^k$  converges to  $x^*$ . Therefore,  $x^*$  is locally stable.  $\square$

The next theorem is easy to prove.

**THEOREM 5.** Let  $x^*$  be a locally stable minimum point of the continuous function  $f(x)$  and let  $g(x)$  be a continuous function. Then for sufficiently small  $\varepsilon > 0$ , the function  $f(x) + \varepsilon g(x)$  has a local minimum point  $x_\varepsilon$  in a neighborhood of  $x^*$  and  $x_\varepsilon \rightarrow x^*$  as  $\varepsilon \rightarrow 0$ .  $\square$

Thus, the stability property implies that the minimum point of the initial function and that of the “perturbed” function are close.

A nonsingular minimum point, as follows from Theorems 2 and 4, is locally stable. In this case, the result of Theorem 5 can be refined.

**THEOREM 6.** Let  $x^*$  be a nonsingular minimum point of  $f(x)$  and let a function  $g(x)$  be continuously differentiable in a neighborhood of  $x^*$ . Then for sufficiently small  $\varepsilon > 0$  there exists a local minimum point  $x_\varepsilon$  of the function  $f(x) + \varepsilon g(x)$  in a neighborhood of  $x^*$ , and

$$x_\varepsilon = x^* - \varepsilon [\nabla^2 f(x^*)]^{-1} \nabla g(x^*) + o(\varepsilon). \quad \square \quad (1)$$

One can also introduce the notion of global stability of minimum points. This can be done by replacing the word “local” by the word “global” in the definition. Namely, a global minimum point is said to be *globally stable* if any minimizing sequence converges to it. In this case we speak of global stability of the minimization problem. Repeating almost verbatim the proof of Theorem 4, we obtain that if  $x^*$  is the unique global minimum point of the continuous function  $f(x)$  and the set  $Q_\alpha = \{x: f(x) \leq \alpha\}$  is nonempty and bounded for some  $\alpha > f(x^*)$ , then  $x^*$  is globally stable. The requirement for the  $Q_\alpha$  to be bounded is essential. For example, for the function  $f(x) = x^2/(1+x^4)$ ,  $x \in \mathbf{R}^1$ , the global minimum point  $x^* = 0$  is unique but not globally stable (since the minimizing sequence  $x^k \rightarrow \infty$  does not converge to  $x^*$ ).

One could introduce the following broader definition of stability which does not include uniqueness of a minimum. The set  $X^*$  of global minimum points of  $f(x)$  is said to be *weakly stable* if all limit points of any minimizing sequence belong to  $X^*$ . A criterion for weak stability is given in Exercise 5.

In addition to a qualitative characteristic (that is, whether a minimum point is stable or not), it is important to have quantitative estimates of stability. Such estimates, which allow one to judge the closeness of  $x$  to a solution  $x^*$  if  $f(x)$  is close to  $f(x^*)$ , have been derived for strongly convex functions. In fact, from (35) of Section 1.1 we have

$$\|x - x^*\|^2 \leq 2\ell^{-1}(f(x) - f(x^*)), \quad (2)$$

where  $\ell$  is the constant of strong convexity. A similar local estimate holds for nonsingular minimum point:

$$\|x - x^*\|^2 \leq 2\ell^{-1}(f(x) - f(x^*)) + o(f(x) - f(x^*)) , \quad (3)$$

where  $\ell$  is the smallest eigenvalue of the matrix  $\nabla^2 f(x^*)$ .

Thus, the number  $\ell$  characterizes the “stability margin” of a minimum point. However,  $\ell$  is not always convenient as a measure of stability—for instance, it varies when  $f(x)$  is multiplied by a constant. Hence the following “normalized” characteristic is often used.

We call the quantity

$$\mu = \overline{\lim_{\delta \rightarrow 0}} \left[ \sup_{x \in L_\delta} \|x - x^*\|^2 / \inf_{x \in L_\delta} \|x - x^*\|^2 \right] , \quad (4)$$

$$L_\delta = \{x: f(x) = f(x^*) + \delta\}$$

$\hookrightarrow$  condition number

the ~~ridge index~~ of a minimum point  $x^*$ . In other words,  $\mu$  characterizes the degree of elongation of the level lines of  $f(x)$  in a neighborhood of  $x^*$ . It is clear that  $\mu \geq 1$ . If  $\mu$  is large, then the level lines are strongly elongated, the function has a gullied character, i.e., it increases sharply in some directions and varies little in other directions. In such cases one speaks of *ill-posed* minimization problems. But if  $\mu$  is close to 1, the level lines of  $f(x)$  are close to being spheres—this corresponds to a well-posed problem. We will see below that the index  $\mu$  is relevant to many problems involving unconstrained minimization and can serve as a measure of the complexity of the problem.

For a quadratic function

$$f(x) = (Ax, x)/2 - (b, x) , \quad A > 0 , \quad (5)$$

we have  $L_\delta = \{x: (A(x-x^*), x-x^*) = 2\delta\}$ . Hence the maximum of  $\|x - x^*\|$  for  $x \in L_\delta$  is attained at  $x_1 = x^* + \gamma_1 \ell_1$ , where  $\ell_1$  is the normalized eigenvector corresponding to the smallest eigenvalue  $\lambda_1$  of the matrix  $A$  and the factor  $\gamma_1$  is determined from the condition  $x_1 \in L_\delta$ , i.e.,  $\lambda_1 \gamma_1^2 = 2\delta$ ,  $\gamma_1 = (2\delta/\lambda_1)^{1/2}$ . Similarly, the minimum of  $\|x - x^*\|$  for  $x \in L_\delta$  is attained on a vector  $x_n = x^* + \gamma_n \ell_n$ ,  $\ell_n$  being the eigenvector corresponding to the largest eigenvalue  $\lambda_n$ ,  $\gamma_n = (2\delta/\lambda_n)^{1/2}$  (Fig. 4). Thus the ratio

$$\mu(\delta) = \|x_1 - x^*\|^2 / \|x_n - x^*\|^2 = \gamma_1^2 / \gamma_n^2 = \lambda_n / \lambda_1$$

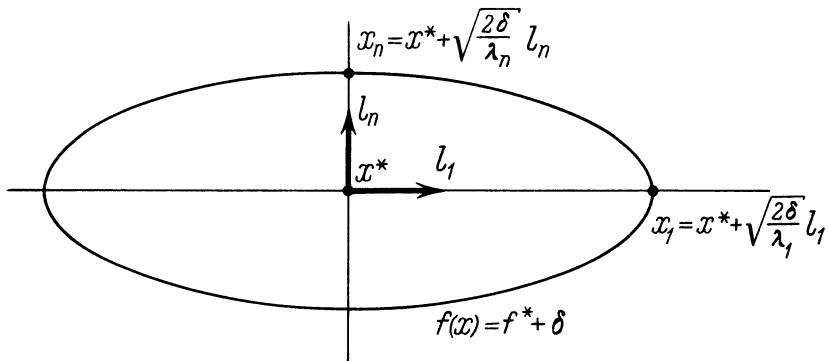


Fig. 4 Condition number of a quadratic function.

$\sqrt{\frac{\lambda_n}{\lambda_1}}$  to the smallest

does not really depend on  $\delta$  and

$$\mu = \frac{\lambda_n}{\lambda_1}. \quad (6)$$

Note that in Linear Algebra the ratio of the largest eigenvalue to the smallest is called the *condition number of the matrix*.

For the case of a nonquadratic function, the condition number of the problem of minimizing the function is equal to the condition number of the Hessian at a minimum point. In fact, if  $x^*$  is a nonsingular minimum point, then

$$\mu = \frac{L}{\ell}, \quad (7)$$

where  $L$  is the largest eigenvalue and  $\ell$  is the smallest eigenvalue of the matrix  $\nabla^2 f(x^*)$ .

We will see later that unstable or ill-posed optimization problems often arise in practical implementation. Methods for solving such problems will be discussed in Section 6.1.

### Exercises

3. Show that a minimum point of a strictly convex continuous function is globally stable.
4. Verify that under the conditions stated in Exercise 2 the set of minimum points is weakly stable.
5. Prove that if  $f(x)$  is continuous and  $Q_\alpha = \{x: f(x) \leq \alpha\}$  is nonempty and bounded for some  $\alpha > \inf f(x)$ , then the set of minimum points of  $f(x)$  is weakly stable.

6. Show that the condition number of a problem does not change under monotone transformations of the function and orthogonal transformations of the variables, i.e., the condition number of  $f(x)$  and  $f_1(x) = \phi(f(Ux))$  are the same if  $\phi: \mathbf{R}^1 \rightarrow \mathbf{R}^1$  is a monotonically increasing continuous function and  $U$  is an orthogonal matrix.
7. Check that for the function  $f(x) = x_1^2 + x_2^4$ , the condition number of a minimum point is infinity.
8. Prove that for a differentiable function  $f(x)$ , the inequality  $f(x) - f(x^*) \geq \alpha \|x - x^*\|$ ,  $\alpha > 0$ , is impossible.

## 1.4 THE GRADIENT METHOD

### 1.4.1 Heuristic Considerations

We now proceed to analyze the methods of unconstrained minimization: the gradient method and Newton's method. These methods, though rarely implemented in "pure form," are models for constructing more realistic algorithms. We will give various proofs of convergence, describe a general technique for constructing proofs, and discuss the theoretical aspects versus the implementation of these methods.

Suppose that at any point  $x$  one can compute the gradient of a function  $\nabla f(x)$ . In this case, the simplest method for minimizing  $f(x)$  is the *gradient* method, in which, starting from some initial approximation  $x^0$ , one constructs an iteration sequence

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k), \quad (1)$$

where the parameter  $\gamma_k \geq 0$  is the step size. Various considerations lead to method (1).

First, recall that in proving necessary conditions for an extremum (Theorem 1 of Section 1.2) we used the fact that if the extremum condition does not hold at  $x$  ( $\nabla f(x) \neq 0$ ), then the value of the function can be decreased by passing to the point  $x - \tau \nabla f(x)$  for a sufficiently small  $\tau > 0$ . By applying this procedure iteratively, we arrive at method (1).

Second, at a point  $x^k$  the differentiable function  $f(x)$  is approximated by the linear function  $f_k(x) = f(x^k) + (\nabla f(x^k), x - x^k)$  to within terms of order  $o(x - x^k)$ . Hence one can seek the minimum of the approximation of  $f_k(x)$  in a neighborhood of  $x^k$ . For example, one can specify an  $\varepsilon_k$  and solve the auxiliary problem

$$\min_{\|x-x^k\| \leq \varepsilon_k} f_k(x). \quad (2)$$

It is natural to adopt its solution as the new approximation  $x^{k+1}$ . One can remain in the neighborhood of  $x^k$  in a different way, too, by adding to  $f_k(x)$  a “penalty” for deviating from  $x^k$ . Thus, one can solve the auxiliary problem

$$\min [f_k(x) + \alpha_k \|x - x^k\|^2] \quad (3)$$

and take its solution as  $x^{k+1}$ . We leave it to the reader to see that a solution of problem (2), (3) is given by formula (1).

Third, at a point  $x^k$  one can choose the direction of *local steepest descent*, i.e., the direction  $y^k$ ,  $\|y^k\| = 1$ , for which the minimum  $f'(x^k; y)$  is attained. Using formula (6) of Section 1.1 for the directional derivative, we obtain

$$y^k = \underset{\|y\|=1}{\operatorname{argmin}} (\nabla f(x^k), y) = -\nabla f(x^k)/\|\nabla f(x^k)\|. \quad (4)$$

Thus, the steepest descent direction is opposite to the gradient direction.

We have examined these arguments so closely because we shall be using them to construct optimization methods in more complex situations (for example, under constraints). However, in such situations these approaches lead to different methods.

### 1.4.2 Convergence

We consider the simplest variant of the gradient method, where  $\gamma_k \equiv \gamma$ :

$$x^{k+1} = x^k - \gamma \nabla f(x^k). \quad (5)$$

We are interested in observing the behavior of this method under various assumptions concerning  $f(x)$  and  $\gamma$ .

**THEOREM 1.** Let  $f(x)$  be differentiable on  $\mathbf{R}^n$ , let the gradient of the  $f(x)$  satisfy a Lipschitz condition:

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|, \quad (6)$$

let the  $f(x)$  be bounded below:

$$f(x) \geq f^* > -\infty, \quad (7)$$

and let  $\gamma$  satisfy the condition

$$0 < \gamma < 2/L. \quad (8)$$

Then, in method (5) the gradient tends to zero:

$$\lim_{k \rightarrow \infty} \nabla f(x^k) = 0,$$

and the function  $f(x)$  monotonically decreases:  $f(x^{k+1}) \leq f(x^k)$ .

**PROOF.** In formula (8) of Section 1.1 we substitute  $x = x^k$ ,  $y = -\gamma \nabla f(x^k)$  and use (6):

$$\begin{aligned} f(x^{k+1}) &= f(x^k) - \gamma \|\nabla f(x^k)\|^2 \\ &\quad - \gamma \int_0^1 (\nabla f(x^k) - \tau \gamma \nabla f(x^k)) - \nabla f(x^k), \nabla f(x^k) d\tau \\ &\leq f(x^k) - \gamma \|\nabla f(x^k)\|^2 + L\gamma^2 \|\nabla f(x^k)\|^2 \int_0^1 \tau d\tau \\ &= f(x^k) - \gamma(1 - \frac{1}{2}L\gamma) \|\nabla f(x^k)\|^2. \end{aligned}$$

Summing the inequalities

$$f(x^{s+1}) \leq f(x^k) - \alpha \|\nabla f(x^k)\|^2, \quad \alpha = \gamma(1 - L\gamma/2) \quad (9)$$

from 0 to  $s$  over  $k$ , we obtain

$$f(x^{s+1}) \leq f(x^0) - \alpha \sum_{k=0}^s \|\nabla f(x^k)\|^2.$$

Since  $\alpha > 0$  by virtue of (8), we have

$$\sum_{k=0}^s \|\nabla f(x^k)\|^2 \leq \alpha^{-1} (f(x^0) - f(x^{s+1})) \leq \alpha^{-1} (f(x^0) - f^*)$$

for all  $s$ , i.e.,  $\sum_{k=0}^{\infty} \|\nabla f(x^k)\|^2 < \infty$ , yielding  $\|\nabla f(x^k)\| \rightarrow 0$ .  $\square$

Now, we show that all the conditions of this theorem are essential. Violations of condition (6) can be twofold:

1) the function  $f(x)$  can be insufficiently smooth at some point. For example, let  $f(x) = \|x\|^{1+\alpha}$ ,  $0 < \alpha < 1$ . This function is differentiable but its gradient does not satisfy a Lipschitz condition since  $\|\nabla f(x) - \nabla f(0)\|/\|x - 0\| = (\alpha + 1)\|x\|^{\alpha-1} \rightarrow \infty$  as  $\|x\| \rightarrow 0$ . In this case one has

$$\gamma \|\nabla f(x^k)\| \gg \|x^k - x^*\| = \|x^k\|$$

for small  $\|x^k\|$ , i.e., the step size in method (5) is large and  $f(x)$  does not decrease monotonically;

2) inequality (6) does not hold for functions that grow faster than a quadratic function. For example, let  $f(x) = \|x\|^{2+\alpha}$ ,  $\alpha > 0$ . Then  $\|\nabla f(x) - \nabla f(0)\|/\|x - 0\| = (2 + \alpha)\|x\|^\alpha \rightarrow \infty$  as  $\|x\| \rightarrow \infty$ . For every  $\gamma > 0$  one can find an  $x^0$  such that method (5), when applied to the function  $\|x\|^{2+\alpha}$ ,  $\alpha > 0$ , with initial approximation  $x^0$ , diverges since one has  $\|x^{k+1}\| > \|x^k\|$ ,  $k = 0, 1, \dots$ .

If condition (7) does not hold, then the function  $f(x)$  does not attain a minimum and the gradient in method (5) does not necessarily tend to zero (for instance, if  $f(x)$  is linear:  $f(x) = (c, x)$ , then  $\|\nabla f(x)\| \equiv \|c\| > 0$ ).

Finally, it is also generally impossible to choose  $\gamma$ , violating condition (8), as is seen from  $f(x) = Lx^2/2$ ,  $x \in \mathbb{R}^1$ . Indeed, if  $\gamma \geq 2/L$ , then in method (5) for this function one has  $f(x^{k+1}) \geq f(x^k)$ ,  $k = 0, 1, \dots$ , for any  $x^0$ .

On the other hand, under the assumptions made in Theorem 1 one cannot prove anything more, viz. the convergence of the sequence  $x^k$ . The function  $f(x) = 1/(1 + \|x\|^2)$  is a good illustration in this case: it satisfies the conditions of the theorem and one has  $\|x^k\| \rightarrow \infty$  for any  $x^0 \neq 0$ .

If we require that  $f(x) \neq f(x^0)$  be bounded, then we can find a subsequence of  $x^k$  converging to some stationary point  $x^*$ . However,  $x^*$  does not need to be a local or a global minimum point. In particular, the gradient method (5) (or even (1) with an arbitrary choice of  $\gamma_k$ ) originated at some stationary point  $x^0$ , remains at this point:  $x^k = x^0$  for all  $k$ . In other words, the gradient method “gets stuck” at any stationary point, whether it is a minimum point, or a saddle point. In finding a global minimum, the gradient method does not “distinguish” local minimum points from global minimum points and there is no guarantee of convergence to a global minimum.

Finally, under the conditions of Theorem 1 the rate of convergence of  $\nabla f(x^k)$  to zero can be very slow. For example, for  $f(x) = 1/x$  for  $x \geq 1$  (the form of  $f(x)$  for  $x < 1$  is immaterial), method (5) for  $\gamma = 1$ ,  $x^0 = 1$  takes the form  $x^{k+1} = x^k + (x^k)^{-2}$ , and one can then show, using Lemma 6 of Section 2.2, that  $|f'(x^k)| = O(k^{-2/3})$ .

Now let us examine the behavior of the gradient method for a narrower class of functions, viz. strongly convex functions, when it is possible to prove stronger results than in Theorem 1: the iterations  $x^k$  converge to a global minimum point with the rate of geometric progression. To this end, we need some inequalities for differentiable, convex and strongly convex functions.

**LEMMA 1.** Let  $f(x)$  be differentiable, let  $\nabla f(x)$  satisfy a Lipschitz condition with constant  $L$  and let  $f(x) \geq f^*$  for all  $x$ . Then

$$\|\nabla f(x)\|^2 \leq 2L(f(x) - f^*) . \quad (10)$$

PROOF. From  $x$  we make a step of the gradient method with  $\gamma \neq 1/L$ . Then (see (9))

$$f^* \leq f(x - L^{-1} \nabla f(x)) \leq f(x) - (2L)^{-1} \|\nabla f(x)\|^2. \quad \square$$

**LEMMA 2.** Let  $f(x)$  be convex and differentiable, and let  $\nabla f(x)$  satisfy a Lipschitz condition with constant  $L$ . Then

$$(\nabla f(x) - \nabla f(y), x - y) \geq L^{-1} \|\nabla f(x) - \nabla f(y)\|^2. \quad (11)$$

**PROOF.** We prove (11) only for twice-differentiable functions. Then (see (13) of Section 1.1)

$$\nabla f(y) = \nabla f(x) + \int_0^1 \nabla^2 f(x + \tau(y-x))(y-x) d\tau = \nabla f(x) + A(y-x),$$

where the matrix

$$A = \int_0^1 \nabla^2 f(x + \tau(y-x)) d\tau$$

is symmetric and nonnegative definite by virtue of (32) of Section 1.1, i.e.,  $A \geq 0$ . Moreover,  $\|A\| \leq L$  since  $\|\nabla^2 f(x)\| \leq L$  for all  $x$  by a Lipschitz condition on the gradient. Hence

$$\begin{aligned} (\nabla f(x) - \nabla f(y), x - y) &= (A(x-y), x - y) \\ &\geq \|A\|^{-1} \|A(x-y)\|^2 \geq L^{-1} \|\nabla f(x) - \nabla f(y)\|^2. \quad \square \end{aligned}$$

**LEMMA 3.** Let  $f(x)$  be a differentiable strongly convex (with constant  $\ell$ ) function and let  $x^*$  be its minimum point (it exists; see Exercise 1 of Section 1.3). Then

$$\|\nabla f(x)\|^2 \geq 2\ell(f(x) - f(x^*)). \quad \square$$

**THEOREM 2.** Let  $f(x)$  be differentiable on  $\mathbf{R}^n$ , let its gradient satisfy a Lipschitz condition with constant  $L$  and let  $f(x)$  be a strongly convex function with constant  $\ell$ . Then for  $0 < \gamma < 2/L$  method (5) converges to a unique global minimum point  $x^*$  with the rate of geometric progression:

$$\|x^k - x^*\| \leq cq^k, \quad 0 \leq q < 1. \quad (12)$$

**PROOF.** All conditions of Theorem 1 are satisfied. Therefore (9) holds:

$$f(x^{k+1}) \leq f(x^k) - \gamma(1 - L\gamma/2) \|\nabla f(x^k)\|^2.$$

We use Lemma 3:

$$f(x^{k+1}) \leq f(x^k) - \ell\gamma(2 - L\gamma)(f(x^k) - f(x^*))$$

yielding

$$\begin{aligned} f(x^{k+1}) - f(x^*) &\leq (1 - \ell\gamma(2 - L\gamma))(f(x^k) - f(x^*)) \\ &= q_1(f(x^k) - f(x^*)) , \end{aligned}$$

$$f(x^k) - f(x^*) \leq q_1^k(f(x^0) - f(x^*)) , \quad q_1 = 1 - 2\ell\gamma + L\ell\gamma^2 .$$

Since  $0 < \gamma < 2/L$ , then  $0 < q_1 < 1$ , and therefore  $f(x^k) \rightarrow f(x^*)$ . From inequality (35) of Section 1.1 we have

$$\|x^k - x^*\|^2 \leq (2/\ell) q_1^k(f(x^0) - f(x^*)) . \quad \square$$

Let us consider an even smaller class of functions—strongly convex twice-differentiable functions.

**THEOREM 3.** Let  $f(x)$  be twice differentiable and let

$$\ell I \leq \nabla^2 f(x) \leq LI , \quad \ell > 0 , \quad (13)$$

for all  $x$ . Then for  $0 < \gamma < 2/L$

$$\|x^k - x^*\| \leq \|x^0 - x^*\| q^k , \quad q = \max \{|1 - \gamma\ell|, |1 - \gamma L|\} < 1 . \quad (14)$$

The quantity  $q$  is minimal and equal to

$$q^* = (L - \ell)/(L + \ell) \quad \text{for } \gamma = \gamma^* = 2/(L + \ell) . \quad (15)$$

**PROOF.** By formula (13) of Section 1.1,

$$\nabla f(x^k) = \nabla f(x^*) + \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))(x^k - x^*) d\tau = A_k(x^k - x^*) ,$$

where  $\ell I \leq A_k \leq LI$  by virtue of (13). Hence

$$\begin{aligned} \|x^{k+1} - x^*\| &= \|x^k - x^* - \gamma \nabla f(x^k)\| = \|(I - \gamma A_k)(x^k - x^*)\| \\ &\leq \|I - \gamma A_k\| \|x^k - x^*\| . \end{aligned}$$

For every symmetric matrix  $A$  we have  $\|I - A\| = \max\{|1 - \lambda_1|, |1 - \lambda_n|\}$ , where  $\lambda_1$  and  $\lambda_n$  are respectively the smallest and the largest eigenvalues of  $A$ . Hence  $\|x^{k+1} - x^*\| \leq q \|x^k - x^*\|$ ,  $q = \max\{|1 - \gamma\ell|, |1 - \gamma L|\}$ . Since  $0 < \gamma < 2/L$ ,  $0 < \ell \leq L$ , then  $|1 - \gamma\ell| < 1$ ,  $|1 - \gamma L| < 1$ , i.e.,  $q < 1$ . Minimizing  $q$  over  $\gamma$ , we obtain (15).  $\square$

We show next that the estimate of the convergence rate given by Theorem 3 is exact and attainable for any quadratic function. Let

$$f(x) = (Ax, x)/2 - (b, x), \quad A > 0, \quad 0 < \ell = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n = L,$$

where the  $\lambda_i$  are the eigenvalues of  $A$ . Take an arbitrary  $0 < \gamma < 2/L$ . Assume that  $|1 - \gamma\ell| \geq |1 - \gamma L|$ . Take  $x^0 = x^* + e^1$ , where  $e^1$  is the normalized eigenvector corresponding to  $\lambda_1$ . Then

$$\begin{aligned} x^k - x^* &= (I - \gamma A)^k (x^0 - x^*) = (1 - \gamma\lambda_1)^k e^1, \\ \|x^k - x^*\| &= |(1 - \gamma\ell)|^k = q^k \|x^0 - x^*\|. \end{aligned}$$

Similarly, if  $|1 - \gamma L| \geq |1 - \gamma\ell|$ , take  $x^0 = x^* + e^n$ , where  $e^n$  is the normalized eigenvector corresponding to  $\lambda_n$ . Then, in the same way,

$$\|x^k - x^*\| = |(1 - \gamma L)|^k = q^k \|x^0 - x^*\|.$$

Therefore, for every  $0 < \gamma < 2/L$ , we can find an  $x^0$  such that  $\|x^k - x^*\| = q^k \|x^0 - x^*\|$ ,  $q = \max\{|1 - \gamma\ell|, |1 - \gamma L|\}$ .

The estimate

$$\|x^k - x^*\| \leq (q^*)^k \|x^0 - x^*\|, \quad q^* = (L - \ell)/(L + \ell)$$

cannot be improved even if  $\gamma$  is optimal for each  $x^0$ . Indeed, take  $x^0 = x^* + e^1 + e^n$  (the notation is the same as above). Then for any  $0 < \gamma < 2/L$ ,

$$x^k - x^* = (I - \gamma A)^k (x^0 - x^*) = (1 - \gamma\ell)^k e^1 + (1 - \gamma L)^k e^n,$$

$$\|x^k - x^*\| = [(1 - \gamma\ell)^{2k} + (1 - \gamma L)^{2k}]^{1/2} \|x^0 - x^*\|/\sqrt{2}.$$

Hence, if either  $|1 - \gamma\ell| > q^*$  or  $|1 - \gamma L| > q^*$ , then  $\|x^k - x^*\|$  decreases slower than  $(q^*)^k$ . But  $q = \max\{|1 - \gamma\ell|, |1 - \gamma L|\} \leq q^*$  only for  $\gamma \neq \gamma^*$ , and

$$|1 - \gamma^*\ell| = |1 - \gamma^*L| = q^* \quad \text{and} \quad \|x^k - x^*\| = (q^*)^k \|x^0 - x^*\|.$$

An analogous argument is valid for any point  $x^0$  such that  $(x^0 - x^*, e^1) \neq 0$ ,  $(x^0 - x^*, e^n) \neq 0$ .

A local analog of Theorem 3 is valid for nonconvex functions as well.

**THEOREM 4.** Let  $x^*$  be a nonsingular local minimum point of  $f(x)$ . Then for  $0 < \gamma < 2/\|\nabla^2 f(x^*)\|$ , method (5) converges locally to  $x^*$  with the rate of geometric progression, i.e., for any  $\delta > 0$  we can find an  $\varepsilon > 0$  such that for  $\|x^0 - x^*\| \leq \varepsilon$ ,

$$\|x^k - x^*\| \leq \|x^0 - x^*\|(q + \delta)^k, \quad (16)$$

$$q = \max \{ |1 - \gamma\ell|, |1 - \gamma L| \} < 1, \quad 0 < \ell I \leq \nabla^2 f(x^*) \leq LI.$$

The quantity  $q$  is minimal and equal to

$$q^* = (L - \ell)/(L + \ell) \quad \text{for } \gamma^* = 2/(L + \ell). \quad \square$$

Other theorems on convergence of gradient methods under somewhat different assumptions will be given in later chapters.

## Exercises

1. Analyze in detail the behavior of the gradient method (5) for the following functions on  $\mathbf{R}^1$ : (a)  $|x|^{1+\alpha}$ ,  $0 < \alpha < 1$ ; (b)  $|x|^{2+\alpha}$ ,  $\alpha > 0$ ; (c)  $x^2$ ; (d)  $(1 + x^2)^{-1}$ . For which  $x^0$  and  $\gamma$  does the method converge? For which does it diverge?

ANSWERS: (a) No convergence for any  $\gamma > 0$  and  $x^0 \neq 0$ , with  $|x^k| \rightarrow [(1/2)(1 + \alpha)\gamma]^{1/(1-\alpha)}$  and the signs of  $x^k$  and  $x^{k+1}$  alternate for  $k \geq k_0$ . (b) The method converges if  $\gamma(2 + \alpha)|x^0|^\alpha \neq 2$  and diverges otherwise, with  $|x^k| \equiv |x^0|$  for  $\gamma(2 + \alpha)|x^0|^\alpha = 2$  and  $|x^k| \rightarrow \infty$  for  $\gamma(2 + \alpha)|x^0|^\alpha > 2$ .

(c) The method converges for  $0 < \gamma < 2$  and diverges for  $\gamma \geq 2$  and any  $x^0 \neq 0$ , with  $|x^k| \equiv |x^0|$  if  $\gamma = 2$  and  $|x^k| \rightarrow \infty$  for  $\gamma > 2$ . (d)  $|x^k| \rightarrow \infty$  for any  $x^0 \neq 0$ .

2. Using the inequality  $\sum_{k=0}^{\infty} \|\nabla f(x^k)\|^2 < \infty$  obtained in proving Theorem 1, show that under the conditions of Theorem 1,

$$\lim_{k \rightarrow \infty} k \|\nabla f(x^k)\|^2 = 0.$$

## 1.5 NEWTON'S METHOD

### 1.5.1 Heuristic Considerations

In the gradient method, the notion of local linear approximation of the objective function  $f(x)$  is basic. If the function is twice differentiable, one

may naturally try to use its quadratic approximation at a point  $x^k$ , i.e., the function

$$f_k(x) = f(x^k) + (\nabla f(x^k), x - x^k) + (\nabla^2 f(x^k)(x - x^k), x - x^k)/2. \quad (1)$$

In the gradient method the next approximation  $x^{k+1}$  was sought under the condition that the linear approximation be a minimum point under the additional constraints of being near to  $x^k$  (since a linear function does not attain its minimum on the entire space): see (2), (3) and (4) of Section 1.4. For a quadratic approximation one can try to impose no restrictions of this kind, since for  $\nabla^2 f(x^k) > 0$  the function  $f_k(x)$  attains an unconstrained minimum. Let us take a minimum point of  $f_k(x)$  as the new approximation:

$$\sqrt{x^{k+1}} = \arg \min_{x \in \mathbb{R}^n} f_k(x).$$

We thus obtain

$$x^{k+1} = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k). \quad (2)$$

One can also arrive at this method, taking a different approach. The minimum point must be a solution of the system of  $n$  equations with  $n$  variables

$$\nabla f(x) = 0. \quad (3)$$

One of the basic methods for solving such systems is Newton's method, which consists in *linearizing* the equations at a point  $x^k$  and solving the linearized system (see Subsection 1.5.3 below). This linearized system in the given case has the form

$$\nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = 0 \quad (4)$$

and its solution  $x^{k+1}$  is given by formula (2).

### 1.5.2 Convergence

**THEOREM 1.** Let  $f(x)$  be twice differentiable, let  $\nabla^2 f(x)$  satisfy a Lipschitz condition with constant  $L$ , let  $f(x)$  be strongly convex with constant  $\ell$ , and let the initial approximation satisfy the condition

$$q = (L\ell^{-2}/2) \|\nabla f(x^0)\| < 1. \quad (5)$$

Then method (2) converges to the global minimum point  $x^*$  with the quadratic rate:

$$\|x^k - x^*\| \leq (2\ell/L)q^{2^k}. \quad (6)$$

**PROOF.** It follows from Lipschitz conditions on  $\nabla^2 f(x)$  that (see (15) of Section 1.1)

$$\|\nabla f(x + y) - \nabla f(x) - \nabla^2 f(x)y\| \leq (L/2)\|y\|^2,$$

where

$$x = x^k, \quad y = -[\nabla^2 f(x^k)]^{-1} \nabla f(x^k).$$

Then  $x + y = x^{k+1}$  and

$$\|\nabla f(x^{k+1})\| \leq (L/2)\|[\nabla^2 f(x^k)]^{-1} \nabla f(x^k)\|^2 \leq (L/2)\|[\nabla^2 f(x^k)]^{-1}\|^2 \|\nabla f(x^k)\|^2.$$

Since  $\nabla^2 f(x^k) \geq \ell I$  (the strong convexity condition, see (33) of Section 1.1), then

$$[\nabla^2 f(x^k)]^{-1} \leq \ell^{-1} I \quad \text{and} \quad \|[\nabla^2 f(x^k)]^{-1}\| \leq \ell^{-1},$$

i.e.,

$$\|\nabla f(x^{k+1})\| \leq (L\ell^{-2}/2)\|\nabla f(x^k)\|^2.$$

Iterating this inequality, we obtain

$$\|\nabla f(x^k)\| \leq \frac{2\ell^2}{L} \left( \frac{L}{2\ell^2} \|\nabla f(x^0)\| \right)^{2^k} = \frac{2\ell^2}{L} q^{2^k}.$$

Applying (37) of Section 1.1 completes the proof.  $\square$

Let us show now that all the conditions of the theorem are essential and that it is generally impossible to strengthen its assertion. Clearly, the existence of a second derivative is required in the formulation of the method, and the strong convexity condition ensures the existence of  $[\nabla^2 f(x^k)]^{-1}$ . Weaker requirements for smoothness (dropping the Lipschitz condition on  $\nabla^2 f(x)$ ) may diminish the convergence rate of the method. For example, let  $f(x) = |x|^{5/2}$ ,  $x \in \mathbf{R}^1$ . Then for  $x > 0$ ,  $f'(x) = (5/2)x^{3/2}$ ,  $f''(x) = (15/4)x^{1/2}$  and  $f''(x)$  does not satisfy the Lipschitz condition. The method takes the form (for  $x^0 > 0$ )

$$x^{k+1} = x^k - (4/15)(x^k)^{-1/2} (5/2)(x^k)^{3/2} = (1/3)x^k,$$

i.e.,  $x^k = (1/3)^k x^0$  and the method converges to  $x^* = 0$  with the rate of geometric progression (rather than quadratically). Finally, it is impossible to assert that the method converges for just any initial approximation (not satisfying (5)). Suppose the problem consists in minimizing the one-dimen-

sional function a derivative of which is shown in Figure 5. This function is twice differentiable, strongly convex (since  $f''(x) \geq 1/2 > 0$  for all  $x$ ),  $f''(x)$  satisfies a Lipschitz condition and  $x^* = 0$ . However, if one starts the iterative process from any point  $x^0$  with  $|x^0| > 1$ , the method does not converge:  $|x^k| \equiv 1$  for all  $k \geq 1$ .

The conditions of Theorem 1 can be somewhat relaxed only in one instance: the local conditions in place of the global ones on  $f(x)$ .

**THEOREM 2.** Let  $f(x)$  be twice differentiable in a neighborhood  $U$  of a non-singular minimum point  $x^*$ , and let  $\nabla^2 f(x)$  satisfy a Lipschitz condition on  $U$ . Then we can find an  $\varepsilon > 0$  such that for  $\|x^0 - x^*\| \leq \varepsilon$ , method (2) converges to  $x^*$  quadratically.  $\square$

For the quadratic function  $f(x) = (Ax, x)/2 - (b, x)$  with  $A > 0$ , Newton's method converges in one step, i.e.,  $x^1 = x^*$  for any  $x^0$ . This is obvious since the approximating function  $f_0(x)$  coincides with  $f(x)$ . The closer  $f(x)$  is to being quadratic, the faster Newton's method converges. Formally, the smaller the  $L$ , the larger (by hypothesis) the domain of convergence defined by (5) and the faster the convergence rate defined by the quantity  $q$ .

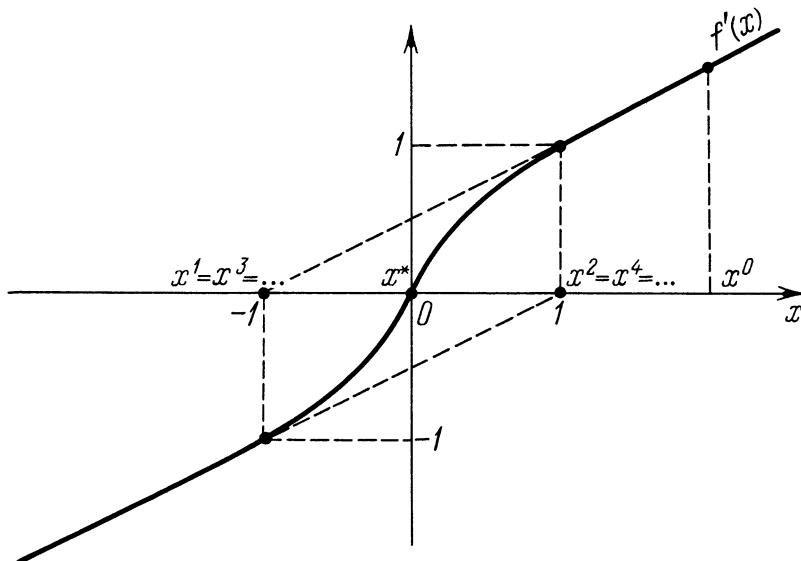


Fig. 5 Divergence of Newton's method.

### 1.5.3 Newton's Method for Solving Equations

Newton's method can be used to solve minimization problems as well as general nonlinear equations:

$$g(x) = 0, \quad g: \mathbf{R}^n \rightarrow \mathbf{R}^n. \quad (7)$$

Newton's method is based on the notion of linear approximation: a linearized equation

$$g(x^k) + g'(x^k)(x - x^k) = 0$$

is solved on the  $k$ th iteration, yielding

$$x^{k+1} = x^k - g'(x^k)^{-1} g(x^k). \quad (8)$$

**THEOREM 3.** Let equation (7) have a solution  $x^*$ , let a function  $g: \mathbf{R}^n \rightarrow \mathbf{R}^n$  be differentiable in a neighborhood of  $x^*$ , and let  $g'(x)$  satisfy a Lipschitz condition in this neighborhood. Furthermore, let the matrix  $g'(x^*)$  be nonsingular. Then we can find an  $\varepsilon > 0$  such that for  $\|x^0 - x^k\| \leq \varepsilon$  method (8) converges to  $x^*$  with the quadratic rate.

It is seen that Theorem 2 is a particular case of Theorem 3 for  $g(x) = \nabla f(x)$ ; the proof of Theorem 3 is the same as that of Theorem 2.  $\square$

We emphasize the fact that for method (8) to converge, we need neither symmetry nor positive definiteness of  $g'(x)$ . In particular, Newton's method is suitable for finding stationary points of a function  $f(x)$  other than minimum points.

## 1.6 THE ROLE OF CONVERGENCE THEOREMS

### 1.6.1 Extreme Viewpoints

Take any book on optimization methods written “by a mathematician for a mathematician”—Cea [0.17] would be a typical example. For the most part it consists of theorems on convergence of the methods. Their formulations are general and abstract, and use the latest machinery of Functional Analysis. The criteria for evaluating the results are the same as in “pure” mathematics: depth, elegance and simplicity of assertions and proofs. Comments and examples are almost totally lacking; a comparative analysis of the methods is absent; there are no numerical examples. The reader who is interested in using the methods has to guess for himself how the mathematical results relate to computational practice, and quite often such a connection

is not simple to establish. It is not rare (especially in periodical literature) to see formal investigation of methods of little interest, including those known to be inefficient. This prompted the publication of a witty parody on “pseudoscientific” works on optimization methods, written by Wolfe [1.11]. Alas, this parody has not remedied the situation—moreover, many readers have taken the article seriously, without comprehending its deliberate absurdity.

Such a situation engendered another extreme approach, which, in essence, rejects the role of theory in the development and study of optimization methods. Its advocates hold the opinion that for creating a method heuristic considerations are quite sufficient. They argue that a rigorous proof of convergence is superfluous, since the conditions of the theorems are hard to check in particular problems, the actual fact of convergence yields little, if anything, and the convergence rate estimates are inaccurate and ineffective. Moreover, in implementing the method, a mass of factors emerge for which a rigorous accounting is impossible (roundoff errors, approximate solution of various auxiliary problems, and so on) and which may strongly affect the course of the implementation process. Therefore, the sole criterion for evaluating a method is how it works out in practice. We shall not elaborate on the subject since this would take us far afield into philosophical questions on the nature of computational mathematics. Rather, with the aid of our results on convergence of the two unconstrained minimization methods, we shall attempt to clarify to what degree convergence theorems can be useful and why they require caution.

### 1.6.2 Why Are Convergence Theorems Necessary?

The answer to this “naive” question is not simple. Of course, for a mathematician dealing with theoretical validation of methods, theorems can be of independent interest in terms of the techniques employed, or the depth of investigation, etc. But how can such theorems be of use to the one who needs to solve a practical problem?

First of all, conditions of the theorems determine the class of problems for which one can count on the applicability of the method. This information is often of a negative nature—if the conditions of the theorem do not hold, then the method may, although not necessarily, be inoperable. Thus, the least restrictive assumptions under which one can prove convergence of the gradient method in the form of (5) of Section 1.4 amount to sufficient smoothness of the objective function (Theorem 1 of Section 1.4). In discussing the theorem, we saw that a violation of these assumptions can indeed make the process diverge. Similarly, in the examples, we saw that stronger smoothness conditions for the functions are also essential for Newton’s method to be implementable. It is convenient when such conditions are of a qualitative nature (smoothness, convexity, and the like), for this allows one to verify them even in complex problems. It is also important that the

conditions in the theorems not be too stringent. For example, as we see from Theorem 3 of Section 1.4, for the gradient method to be used it is necessary that the second derivative exists. However, this condition is superfluous (see Theorem 1 of Section 1.4); one needs it only to estimate the convergence rate. That is why it is useful to have several theorems with assertions concerning the same method but under different assumptions (such as Theorems 1-4 of Section 1.4 for the gradient method).

Also, convergence theorems provide important information on the qualitative behavior of the method: whether it converges for any initial approximation or only for a sufficiently good one, and in what sense it converges (the function converges, or the argument converges, or in the limit, and so on). Thus Theorem 1 of Section 1.4 ensures that the gradient method is applicable from any initial point, yet we assert only that  $\nabla f(x^k) \rightarrow 0$  (while there may be no convergence with respect to the function or argument, as illustrated by the examples). In Theorem 1 of Section 1.5, conversely, convergence of Newton's method (in the argument to a global minimum) is demonstrated only for a good initial approximation and, as we say above, this condition is essential. Therefore, for the implementation of Newton's method one needs to have a good initial approximation; or otherwise, the method may diverge.

The actual proofs of convergence theorems often contain useful information. Most frequently, they are based on the idea that some scalar monotonically decreases in the iterative process (this will be examined in detail in Chapter 2). In Theorems 1 and 2 of Section 1.4 it is the function being minimized; in Theorems 3 and 4 therein it is the distance to the minimum point; and in Theorem 1 of Section 1.5 it is the norm of the gradient. This is often accessible ( $f(x)$ ,  $\|\nabla f(x)\|$ ) and its behavior in the computational process determines the convergence or divergence of the method—if the course of the process is normal, it ought to decrease. If the proof is based, for instance, on monotonic decrease of  $\|x^k - x^*\|$ , it would be unreasonable to require  $f(x)$  be monotonically decreasing at each step.

An estimate of the convergence rate provides especially important information. This information can be of a positive as well as of a negative nature. For example, the estimate of the convergence rate of Newton's method in Theorem 1 of Section 1.5 shows that the method converges very rapidly. Indeed, if the initial approximation is sufficiently close to the solution ( $q < 1$ ), then, according to (6) of Section 1.5,  $\|x^k - x^*\| \leq 2q^{2^k}$  (since  $L > L'$ ). Hence for  $q = 0.5$ , we have  $\|x^k - x^*\| \leq 2^{-2^{k+1}}$ , so that  $\|x^5 - x^*\| < 10^{-9}$ , whereas for  $q = 0.1$ , we have  $\|x^k - x^*\| \leq 2 \cdot 10^{-2^k}$ , so that  $\|x^4 - x^*\| < 10^{-16}$ . In other words, if Newton's method is applicable, no more than four or five iterations are required to obtain a solution with very high accuracy. On the other hand, the gradient method for an optimal choice of  $\gamma$ , by virtue of Theorem 3 of Section 1.4, converges geometrically with ratio  $q = (L-L')/(L+L')$ , and we saw that this estimate was exact for the qua-

dramatic function. For large condition numbers  $\mu = L/\ell$ , the progression ratio  $q \approx 1 - 2/\mu$  close to 1. As we shall see in later chapters, it is not uncommon for very simple problems of mean square approximation by polynomials that  $\mu$  attains values of the order  $10^8$ . Clearly, for  $\mu = 10^8$ , roughly  $5 \cdot 10^7$  iterations are needed to diminish  $\|x^0 - x^*\|$  by a factor of  $e$ . In other words, the gradient method is unfeasible in such a situation. This negative result concerning the behavior of the gradient method can be derived purely theoretically, without any numerical experiments. In comparison with other minimization problems, this is reason enough for adopting a careful attitude towards the gradient method—one can hardly count on this method as an efficient means of solving complex problems.

A theoretical estimation of the convergence rate also shows what exactly determines the behavior of the method. Thus, for the gradient method, “difficult” problems are the ill-posed ones, and the choice of an initial approximation has no influence on the convergence rate; whereas for Newton’s method the rate depends on the quality of the initial approximation as well as the closeness of the function to a quadratic one, but not on the condition number of the problem. For the conjugate-gradient method, as will be seen in the sequel, the dimension of the problem is most crucial in the estimation, in contrast to the gradient method and Newton’s method, one can make an “educated guess” as to a particular method to be used in a specific problem.

Finally, using results on the convergence rate, one can choose in advance (or estimate) the required number of iterations, to achieve the specified accuracy. Thus, if we apply the conditions of Theorem 3 of Section 1.4 and know estimates for  $\ell$ ,  $L$  and  $\|x^0 - x^*\|$ , we can ascertain the number of steps  $k$  yielding the accuracy  $\|x^k - x^*\| \leq \varepsilon$  in the gradient method with the optimal  $\gamma = 2/(L + \ell)$

$$k = \log \frac{\varepsilon}{\|x^0 - x^*\|} / \log \frac{\mu-1}{\mu+1} \approx \frac{\mu}{2} \log \frac{\|x^0 - x^*\|}{\varepsilon}, \quad \mu = \frac{L}{\ell}.$$

### 1.6.3 Proceed With Caution

Let us lend an ear to the criticism of the theoretical approach to studying optimization methods. Advocates of this viewpoint regard the theory relating to this matter as a superfluous and even, sometimes, harmful luxury. They assert that the fact that the method is convergent does not mean that this method is efficient. This is undoubtedly true. Indeed, it is wrong to assume that a given method is to be implemented if its convergence is proved—for the rate of convergence may be hopelessly slow. However, we have remarked above that convergence theorems, including those without estimates of the rate of convergence, provide important information relating to the

range of applicability of the method, its performance, etc. All this information is still not enough to draw definitive conclusions as to whether the method is appropriate and advantageous for solving a particular problem.

Furthermore, results related to the convergence of the method are often doubtful because the assumptions may be difficult to verify, or the parameters are unknown, or the estimates are asymptotic—indeed, such criticism is, to a great extent, justified. Convergence theorems are frequently cumbersome and it is impossible to verify them for some specific problem. The situation gets worse if the assertions are of an *a posteriori* nature—“... suppose that in an iterative process such-and-such a condition holds ...” Why does not one assume simply that  $x^k \rightarrow x^*$ ? Still, the picture is not always so gloomy. As is evident from the theorems of Sections 1.4 and 1.5, the assumptions are simple and general—they require smoothness, convexity, strong convexity, nonsingularity and other similar natural and easily verifiable conditions. The constants  $L$ ,  $\gamma$  and  $q$  in those theorems are indeed usually unknown and therefore a constructive choice of  $\gamma$  in the gradient method or explicit estimates of the rate of convergence are impossible. There are however more complicated ways of choosing  $\gamma_k$  in the gradient method (Chapter 3), based on the theorems of Section 1.4. Although a quantitative estimation of the rate of convergence is not always possible, its qualitative characteristic leaves no doubt. Finally, the estimates of the convergence rate do not have to be asymptotic—in Theorems 2 and 3 of Section 1.4 and Theorem 1 of Section 1.5 they are true for all finite  $k$ .

Yet another drawback of convergence theorems is that they deal with ideal, unrealistic, situations, devoid of noise problems, roundoff errors, unfeasibility of an exact solution of the auxiliary problems, etc., whereas in fact all these factors strongly influence the behavior of the method in the practical implementation. Note that in all of the theorems given above we assumed that the gradient was computed exactly, that inversion of the matrix in Newton's method was error-free, and so on. In Chapter 4 we will discuss these same methods, taking into account noise of different kinds. It is clear that noise ultimately limits efficiency. Hence a comparative evaluation of methods will need to rest on more general convergence theorems which take noise into account.



## CHAPTER 2

### GENERAL SCHEMES FOR INVESTIGATING ITERATIVE METHODS

The results concerning convergence and rates of convergence of minimization algorithms were derived in Chapter 1 without invoking any general theorems. That approach was natural, since the proofs were very simple. However, as the problems and methods get more complex, proving them becomes more cumbersome and more laborious. A close analysis of the proofs shows that the ideas on which the proofs are based are simple and uniform. It is appropriate to put these ideas in explicit form, derive general results and then use them systematically to prove particular algorithms. This is what we shall do in this chapter.

#### 2.1 LYAPUNOV'S FIRST METHOD

Lyapunov's first method consists in linearizing the iterative procedure and evaluating convergence on the basis of the linearized process. But first we recall some essentials of Linear Algebra.

##### 2.1.1 Review of Linear Algebra

Let  $A$  be a square  $n \times n$ -matrix and let  $\lambda_1, \dots, \lambda_n$  be its eigenvalues. By the spectral radius of  $A$  we mean the quantity

$$\rho(A) = \max_{1 \leq i \leq n} |\lambda_i|. \quad (1)$$

$\| \cdot \|$  is another important characteristic of any (may be nonquadratic) matrix. By us-

### 38 Chapter 2 Iterative Methods: General Schemes

The norm

$$\| A \| = \max_{\| x \| = 1} \| Ax \| \quad (2)$$

ing the fact that for a symmetric matrix all the eigenvalues are real-valued and there exists a complete orthogonal system of eigenvectors, it is not hard to prove that  $\rho(A) = \| A \|$  for a symmetric matrix. For a nonsymmetric matrix,  $\rho(A) \leq \| A \|$  and generally  $\rho(A) \neq \| A \|$ . For example, for the matrix  $A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ , both eigenvalues are equal to 0. Hence  $\rho(A) = 0$  but  $\| A \| = 1$ . An important relationship between  $\| A \|$  and  $\rho(A)$  is given by the equality

$$\rho(A) = \lim_{k \rightarrow \infty} \| A^k \|^{1/k}, \quad (3)$$

which implies the following lemma.

**LEMMA 1.** For  $\lim_{k \rightarrow \infty} A^k = 0$  it is necessary and sufficient that  $\rho(A) < 1$  and for every  $\varepsilon > 0$  there be a  $c = c(\varepsilon)$  such that  $\| A^k \| \leq c(\rho(A) + \varepsilon)^k$  for all integers  $k$ .  $\square$

**COROLLARY.** In order that the iterative sequence of vectors  $x^{k+1} = Ax^k$  converge to 0 as  $k \rightarrow \infty$  for any  $x^0$ , it is necessary and sufficient that  $\rho(A) < 1$ .  $\square$

**LEMMA 2.** Let  $\rho(A) < 1$ . Then the matrix equation

$$A^T U A = U - C \quad (4)$$

has a solution  $U$  which is symmetric if  $C$  is symmetric, and  $U \geq C$  if  $C \geq 0$ .

**PROOF.** Since  $\| A^k \| \leq cq^k$ ,  $q < 1$  (Lemma 1), the series  $\sum_{k=0}^{\infty} (A^T)^k C A^k$  converges to some matrix  $U$ . This matrix  $U$  is symmetric if  $C$  is symmetric,  $U \geq 0$  for  $C \geq 0$ ,

$$A^T U A = \sum_{k=1}^{\infty} (A^T)^k C A^k = U - C, \quad U = C + A^T U A \geq C$$

if  $C \geq 0$ .  $\square$

We say that a square matrix  $A$  with eigenvalues  $\lambda_1, \dots, \lambda_n$  is *stable* (or *Hurwicz*) if

$$\operatorname{Re} \lambda_i < 0, \quad i = 1, \dots, n. \quad (5)$$

**LEMMA 3.** For  $\lim_{t \rightarrow \infty} e^{At} = 0$  it is necessary and sufficient that  $A$  be stable. In this case, for every  $\varepsilon > 0$  we can find a  $c = c(\varepsilon)$  such that  $\| e^{At} \| \leq c(\varepsilon)e^{(\gamma+\varepsilon)t}$  for all  $t \geq 0$ ,  $\gamma = \max_i \operatorname{Re} \lambda_i$ .

Indeed, the eigenvalues of  $B = e^A$  are  $e^{\lambda_i}$ , hence  $\rho(B) = \max e^{\operatorname{Re} \lambda_i} = e^\gamma$ . Since  $e^\gamma < 1$  iff  $\gamma < 0$ , then the condition  $\rho(B) < 1$  is equivalent to  $\gamma < 0$ . Now we need to use Lemma 1 (more precisely, its generalization from Exercise 3 below).  $\square$

**LEMMA 4** (Lyapunov). Let the matrix  $A$  be stable and let the matrix  $C$  be symmetric. Then the equation

$$AU + UA^T = -C \quad (6)$$

has a solution, and  $U > 0$  ( $U \geq 0$ ) if  $C > 0$  ( $C \geq 0$ ).

**PROOF.** According to Lemma 3, the matrix  $U = \int_0^\infty e^{At} C e^{A^T t} dt$  is defined. The matrix  $Z(t) = e^{At} C e^{A^T t}$  is a solution of the differential equation  $\dot{Z}(t) = AZ + ZA^T$ ,  $Z(0) = C$ , i.e.,  $U = \int_0^\infty Z(t) dt$ . Hence

$$AU + UA^T = \int_0^\infty (AZ + ZA^T) dt = \int_0^\infty \dot{Z}(t) dt = -Z(0) = -C.$$

Then  $U = \int_0^\infty e^{At} C e^{A^T t} dt$  is the required solution and, also,  $U > 0$  ( $U \geq 0$ ) if  $C > 0$  ( $C \geq 0$ ).  $\square$

The relationship between stable matrices and matrices with  $\rho(A) < 1$  is given by the next lemma.

**LEMMA 5.** Let  $A$  be stable,

$$B = I + \gamma A, \quad 0 < \gamma < \min_i (-2 \operatorname{Re} \lambda_i |\lambda_i|^{-2}).$$

Then  $\rho(B) < 1$ .

Indeed, if  $\lambda_i$  are the eigenvalues of  $A$  and  $\mu_i$  are the eigenvalues of  $B$ , then

$$\mu_i = 1 + \gamma \lambda_i,$$

$$|\mu_i|^2 = (1 + \gamma \operatorname{Re} \lambda_i)^2 + \gamma^2 (\operatorname{Im} \lambda_i)^2 = 1 + 2\gamma \operatorname{Re} \lambda_i + \gamma^2 |\lambda_i|^2 < 1,$$

i.e.,  $\rho(B) < 1$ .  $\square$

## Exercises

1. Show that if the matrix  $A$  is symmetric or has pairwise distinct eigenvalues, then in Lemma 1 one can take  $\varepsilon = 0$ ,  $c(\varepsilon) = 1$ .

2. Given an example of a matrix  $A$  with  $\rho(A) \geq 1$  and some  $x^0 \neq 0$  such that  $A^k x^0 \rightarrow 0$  as  $k \rightarrow \infty$ .
3. Show that Lemma 1 is also valid for nonintegral exponents, i.e.,  $\|A^t\| \leq c(\varepsilon)(\rho(A) + \varepsilon)^t$  for all real  $t \geq 0$ .

### 2.1.2 Theorems on Linear Convergence

We will often use the term *linear convergence* as a synonym for convergence with the rate of geometric progression. Similarly, superlinear convergence stands for convergence more rapid than that defined by any geometric progression. Finally, the term quadratic convergence is used for processes involving an estimate of the form  $u_{k+1} \leq cu_k^2$ , where  $u_k$  is some measure of closeness to the solution in the  $k$ th iteration.

Consider an iterative process of the form

$$x^{k+1} = g(x^k), \quad (7)$$

where  $g$  is some mapping from  $\mathbf{R}^n$  into  $\mathbf{R}^n$ . We call the point  $x^*$  a fixed point for (7), if  $x^* = g(x^*)$ . In this case, for  $x^k = x^*$  one has  $x^s \equiv x^*$  for all  $s \geq k$ .

**THEOREM 1.** Let  $x^*$  be a fixed point of (7), let  $g(x)$  be differentiable and let the spectral radius of the Jacobian  $g'(x^*)$  satisfy the condition  $\rho(g'(x^*)) < 1$ . Then the process (7) converges locally linearly to  $x^*$  and for every  $0 < \varepsilon < 1 - \rho$  we can find a  $\delta > 0$  and a  $c$  such that for all  $k \geq 0$

$$\|x^k - x^*\| \leq c(\rho + \varepsilon)^k \quad (8)$$

for  $\|x^0 - x^*\| \leq \delta$ .

Let us sketch the proof. Let  $A = g'(x^*)$ . Then, by the definition of a derivative,

$$g(x) = g(x^*) + A(x - x^*) + o(x - x^*).$$

Hence (7) can be written in the form

$$z^{k+1} = Az^k + y^k, \quad z^k = x^k - x^*, \quad y^k = o(z^k),$$

implying

$$\begin{aligned} z^{k+1} &= A^{k+1}z^0 + \sum_{i=1}^k A^{k-i}y^i, \\ \|z^{k+1}\| &\leq \|A^{k+1}\| \|z^0\| + \sum_{i=0}^k \|A^{k-i}\| \|y^i\|. \end{aligned} \quad (9)$$

From Lemma 1,  $\|A^k\| \leq c(\varepsilon)(\rho + \varepsilon)^k$ . Substituting the latter into (9) and using the fact that  $\|y^k\| = o(z^k)$  proves the theorem.  $\square$

Theorem 1 guarantees the local convergence of method (7). In certain cases, one can also assert global convergence. One such case is obvious—that of a linear function  $g(x)$ . We give also a result on global convergence for nonlinear functions. We need to consider the iterative process written in the form

$$x^{k+1} = x^k - \gamma(Ax^k + \phi(x^k)). \quad (10)$$

**THEOREM 2.** Let the matrix  $A$  be stable and let  $\phi: \mathbf{R}^n \rightarrow \mathbf{R}^n$  satisfy the condition

$$\|\phi(x)\| \leq L\|x\|.$$

Then, if

$$L < \frac{1}{2\|U\|}, \quad 0 < \gamma < \frac{\|U\|^{-1} - 2L}{(L + \|A\|)^2}, \quad (11)$$

where  $U$  is the solution of the matrix equation

$$UA + A^T U = I, \quad (12)$$

the process (10) converges to zero with the rate of geometric progression for any  $x^0$ :

$$\begin{aligned} \|x^k\|^2 &\leq \|x^0\|^2 \|U^{-1}\| \|U\| q^k, \\ q &= 1 - \left(\frac{1}{2}\right) \gamma \|U\|^{-1} + \gamma L + \left(\frac{1}{2}\right) \gamma^2 (\|A\| + L)^2. \end{aligned} \quad (13)$$

To prove the theorem, it suffices to introduce  $u_k = (Ux^k, x^k)$  and derive the relation  $u_{k+1} \leq qu_k$ .  $\square$

The results obtained above can be used to investigate the finite-difference equations

$$y_k = a_1 y_{k-1} + a_2 y_{k-2} + \cdots + a_n y_{k-n} + \phi(y_{k-1}, \dots, y_{k-n}), \quad (14)$$

where  $y_i \in \mathbf{R}^1$ . For this we introduce the vectors

$$x_k = (y_{k-1}, \dots, y_{k-n}) \in \mathbf{R}^n, \quad x^{k+1} = (y_k, y_{k-1}, \dots, y_{k-n+1}) \in \mathbf{R}^n.$$

Then  $x^{k+1} = Ax^k + h(x^k)$ , where

$$A = \begin{bmatrix} a_1 & a_2 & \dots & a_n \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 \end{bmatrix}, \quad h(x) = \begin{bmatrix} \phi(x) \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \quad (15)$$

Thus the iterative process has been reduced to (7).

This procedure is typical for the investigation of multistep iterative processes in which each approximation depends on several previous approximations. Increasing the dimension of the problem allows reduction to a one-step process.

### Exercise

4

4. Prove that if all the roots of the characteristic equation  $\lambda^n = a_1\lambda^{n-1} + \cdots + a_n$  are of modulus less than 1, then for a matrix  $A$  of the form (15) one has  $\rho(A) < 1$ .

### 2.1.3 A Theorem on Superlinear Convergence

For  $g'(x^*) = 0$ , it follows from Theorem 1 that method (7) converges more rapidly than any geometric progression. This result can be refined.

**THEOREM 3.** Let  $x^*$  be a fixed point of (7), let  $g(x)$  be differentiable in  $S = \{x: \|x - x^*\| \leq \|x^0 - x^*\|\}$ , let  $g'(x)$  satisfy a Lipschitz condition on  $S$ , and let  $g'(x^*) = 0$ . Then, if

$$q = (L/2)\|x^0 - x^*\| < 1, \quad (16)$$

then

$$\|x^k - x^*\| \leq (2/L)q^{2^k}. \quad (17)$$

**PROOF.** Obviously,  $x^0 \in S$ . By formula (15) of Chapter 1, we have

$$\begin{aligned} \|x^1 - x^*\| &= \|g(x^0) - g(x^*) - g'(x^*)(x^0 - x^*)\| \\ &\leq (L/2)\|x^0 - x^*\|^2 \leq q\|x^0 - x^*\|. \end{aligned}$$

Therefore,  $x^1 \in S$ . Similarly,  $x^k \in S$  for all  $k$ . Hence we can use the same estimate:

$$\|x^{k+1} - x^*\| = \|g(x^k) - g(x^*) - g'(x^*)(x^k - x^*)\| \leq (L/2)\|x^k - x^*\|^2. \quad \square$$

### Exercise

5. Let  $x^*$  be a nonsingular minimum point of  $f(x)$  and let  $\nabla^2 f(x)$  satisfy a Lipschitz condition in a neighborhood of  $x^*$ . Then the method

$$x^{k+1} = x^k - [\nabla^2 f(x^*)]^{-1} \nabla f(x^k) \quad (18)$$

converges locally to  $x^*$  with the quadratic rate. Employ Theorem 3 to prove this.

## 2.2 LYAPUNOV'S SECOND METHOD

This is the most commonly used method for proving convergence of iterative processes. The idea is to introduce a certain nonnegative scalar function  $V(x)$  (the Lyapunov function) and examine its values on the sequential iterations  $x^k$ . If the values decrease monotonically and are bounded below, then  $V(x^k) - V(x^{k+1}) \rightarrow 0$ . This, under certain additional assumptions, yields convergence of the method.

If we review the above results from this viewpoint, we see that most of them are derived via this approach. Thus, in proving the gradient method in Chapter 1, the objective function proper,  $f(x) - f^*$ , was a Lyapunov function in Theorem 1 and 2 of Section 1.4 and in Theorems 3 and 4 of Section 1.4 it was the distance to the minimum point. In proving Newton's method (Theorem 1 of Section 1.5), a monotone decrease of the gradient norm was used (that is, the deviation from zero). Finally, in proving Theorem 2 of Section 2.1, a special quadratic Lyapunov function was constructed. Similar procedures of choosing Lyapunov functions are common for other, more complex problems.

### 2.2.1 Lemmas on Numerical Sequences

For values of the Lyapunov function  $u_k = V(x^k)$ , an iteration relation of the form

$$u_{k+1} \leq \phi_k(u_k) \quad (1)$$

holds at the  $k$ th step of the process. Hence the conclusion that  $u_k \rightarrow 0$  and the estimate of the rate of convergence of  $u_k$ . The behavior of sequences of the form (1) for certain “typical” functions  $\phi_k$  is of significance. For example, we have come across some simple relations (1). Say, in proving the convergence of the gradient method (Sec. 1.4), we obtained

$$u_{k+1} \leq q u_k, \quad 0 \leq q < 1, \quad (2)$$

where  $u_k = f(x^k) - f^*$ , or  $u_k = \|x^k - x^*\|^2$ , or  $u_k = \|\nabla f(x^k)\|$ . The estimate  $u_k \leq u_0 q^k$  follows from (2). In proving Newton's method (Sec. 1.5), we obtained for  $u_k = \|\nabla f(x^k)\|$ :

$$u_{k+1} \leq c u_k^2, \quad c > 0, \quad (3)$$

yielding  $u_k \leq c^{-1} (c u_0)^{2^k}$  and if  $c u_0 < 1$  then  $u_k \rightarrow 0$ .

In other problems, relation (1) is more complex and the analysis is not quite so trivial.

We start with linear inequalities of the form

$$u_{k+1} \leq q_k u_k + \alpha_k, \quad q_k \geq 0, \quad (4)$$

implying

$$u_k \leq q_{k-1} q_{k-2} \cdots q_0 u_0 + q_{k-1} \cdots q_1 \alpha_0 + \cdots + q_{k-1} \alpha_{k-2} + \alpha_{k-1}. \quad (5)$$

Now we consider some special cases.

**LEMMA 1.** Let

$$u_{k+1} \leq q u_k + \alpha, \quad 0 \leq q < 1, \quad \alpha > 0. \quad (6)$$

Then

$$u_k \leq \alpha/(1-q) + (u_0 - \alpha/(1-q))q^k. \quad (7)$$

**PROOF.** Setting  $v_k = u_k - \alpha/(1-q)$ , we obtain from (6) that  $v_{k+1} \leq v_k q$ , and therefore (7).  $\square$

Thus,  $u_k$  converges geometrically into the region  $u \leq \alpha/(1-q)$  with ratio  $q$ .

**LEMMA 2.** Let  $u_k \geq 0$  and let

$$u_{k+1} \leq (1 + \alpha_k)u_k + \beta_k, \quad \alpha_k \geq 0, \quad \beta_k \geq 0,$$

$$\sum_{k=0}^{\infty} \alpha_k < \infty, \quad \sum_{k=0}^{\infty} \beta_k < \infty. \quad (8)$$

Then  $u_k \rightarrow u \geq 0$ .

The proof is the same as that of the more general Lemma 9 below.  $\square$

**LEMMA 3.** Let

$$u_{k+1} \leq q_k u_k + \alpha_k, \quad 0 \leq q_k < 1, \quad \alpha_k \geq 0,$$

$$\sum_{k=0}^{\infty} (1 - q_k) = \infty, \quad \alpha_k / (1 - q_k) \rightarrow 0. \quad (9)$$

Then  $\overline{\lim}_{k \rightarrow \infty} u_k \leq 0$ . In particular, if  $u_k \not\downarrow 0$ , then  $u_k \rightarrow 0$ .  $\square$

**COROLLARY.** If in (9)  $q_k \equiv q < 1$ ,  $\alpha_k \rightarrow 0$ ,  $u_k \geq 0$ , then  $u_k \rightarrow 0$ .  $\square$

Under the conditions of Lemma 3, one can also estimate the rate of convergence for a number of cases.

**LEMMA 4** (Chung). Let  $u_k \geq 0$  and

$$u_{k+1} \leq \left(1 - \frac{c}{k}\right) u_k + \frac{d}{k^{p+1}}, \quad d > 0, \quad p > 0, \quad c > 0. \quad (10)$$

Then

$$u_k \leq d(c-p)^{-1} k^{-p} + o(k^{-p}) \quad \text{for } c > p, \quad (11)$$

$$u_k = O(k^{-c} \log k) \quad \text{for } p = c, \quad (12)$$

$$u_k = O(k^{-c}) \quad \text{for } p > c. \quad (13)$$

**PROOF.** For any relation between  $c$  and  $p$  we have that Lemma 3 is applicable since

$$1 - q_k = c/k, \quad \sum_{k=0}^{\infty} (1 - q_k) = \infty, \quad \alpha_k (1 - q_k)^{-1} = dc^{-1} k^{-p} \rightarrow 0,$$

and hence  $u_k \rightarrow 0$ . Let  $c > p$ . Also, let  $v_k = k^p u_k - d(c-p)^{-1}$ . Then

$$\begin{aligned} v_{k+1} &= (k+1)^p u_{k+1} - \frac{d}{c-p} \leq k^p \left(1 + \frac{1}{k}\right)^p \left[ \left(1 - \frac{c}{k}\right) u_k + \frac{d}{k^{p+1}} \right] - \frac{d}{c-p} \\ &= k^p u_k \left(1 - \frac{c-p}{k} + o\left(\frac{1}{k}\right)\right) + \frac{d}{k} \left(1 + \frac{p}{k} + o\left(\frac{1}{k}\right)\right) - \frac{d}{c-p} \\ &= \left(v_k + \frac{d}{c-p}\right) \left(1 - \frac{c-p}{k} + o\left(\frac{1}{k}\right)\right) + \frac{d}{k} \left(1 + \frac{p}{k} + o\left(\frac{1}{k}\right)\right) - \frac{d}{c-p} \\ &= v_k \left(1 - \frac{c-p}{k} + o\left(\frac{1}{k}\right)\right) + \frac{dp}{k^2} + o\left(\frac{1}{k^2}\right). \end{aligned}$$

Applying Lemma 3, we have  $\lim_{k \rightarrow \infty} v_k \leq 0$ , which proves (11).

Now let  $p \geq c$ . Also, let  $v_k = u_k k^c$ . Then

$$\begin{aligned} v_{k+1} &= u_{k+1} (k+1)^c \leq \left[ \left(1 - \frac{c}{k}\right) u_k + \frac{d}{k^{p+1}} \right] k^c \left(1 + \frac{c}{k} + \frac{c^2}{2k^2} + o\left(\frac{1}{k^2}\right)\right) \\ &= \left(1 - \frac{c^2}{2k^2} + o\left(\frac{1}{k^2}\right)\right) v_k + \frac{d}{k^{p-c+1}} \left(1 + O\left(\frac{1}{k}\right)\right) \leq v_k + \frac{d'}{k^{p-c+1}} \end{aligned}$$

for sufficiently large  $k$ . Summing over  $k$ , we obtain that  $v_k$  is bounded for  $p > c$  (since the series  $\sum_{k=1}^{\infty} (1/k^{\alpha})$  converges for  $\alpha > 1$ ) and  $v_k = O(\log k)$  for  $p = c$  (since  $\sum_{i=1}^k (1/i) = O(\log k)$ ). This proves (12) and (13).  $\square$

**LEMMA 5** (Chung). Let  $u_k \geq 0$

$$u_{k+1} \leq \left(1 - \frac{c}{k^s}\right) u_k + \frac{d}{k^t}, \quad 0 < s < 1, \quad s < t. \quad (14)$$

Then

$$u_k \leq \frac{d}{c} \frac{1}{k^{t-s}} + o\left(\frac{1}{k^{t-s}}\right). \quad \square$$

We proceed to investigate recurrence inequalities defined by nonlinear relations.

**LEMMA 6.** Let  $u_k > 0$  and let

$$u_{k+1} \leq u_k - \alpha_k u_k^{1+p}, \quad \alpha_k \geq 0, \quad p > 0. \quad (15)$$

Then

$$u_k \leq u_0 \left(1 + p u_0^p \sum_{i=0}^{k-1} \alpha_i\right)^{-1/p}. \quad (16)$$

In particular, if  $\alpha_k \equiv \alpha$ ,  $p = 1$ , then

$$u_k \leq u_0 / (1 + \alpha k u_0). \quad (17)$$

**PROOF.** We have

$$\begin{aligned} 0 < u_{k+1} &\leq u_k (1 - \alpha_k u_k^p), \quad 1 - \alpha_k u_k^p > 0, \\ \overbrace{u_{k+1}}^p &\geq u_k^{-p} (1 - \alpha_k u_k^p)^{-p} \geq u_k^{-p} (1 + p \alpha_k u_k^p) = u_k^{-p} + p \alpha_k. \end{aligned}$$

$\square$

We use the inequality  $(1-x)^{-p} \geq 1+px$ , which holds for  $x < 1$ ,  $p > 0$ . Summing the inequalities yields (16).  $\square$

**LEMMA 6'.** Let  $u_k \geq 0$  and let

$$u_{k+1} \leq (1+\alpha_k)u_k - \gamma_k\phi(u_k) + \beta_k, \quad \alpha_k \geq 0, \quad \gamma_k \geq 0, \quad \beta_k \geq 0,$$

$$\sum_{k=0}^{\infty} \gamma_k = \infty, \quad \alpha_k \rightarrow 0, \quad \beta_k \rightarrow 0, \quad \frac{\alpha_k}{\gamma_k} \rightarrow 0, \quad \frac{\beta_k}{\gamma_k} \rightarrow 0,$$

$\phi(u) > 0$  for  $u > 0$ ,  $\phi(0) = 0$ ,  $\phi(u') \geq \phi(u)$  for  $u' \geq u \geq 0$ . Also, let either  $\alpha_k \equiv 0$  or  $\phi(u)$  be a convex function. Then  $u_k \rightarrow 0$ .

**PROOF.** Choose  $0 < \theta < 1$  and denote

$$I_1 = \{k: \beta_k > \theta\gamma_k\phi(u_k) - \alpha_k u_k\}, \quad I_2 = \{k: \beta_k \leq \theta\gamma_k\phi(u_k) - \alpha_k u_k\}.$$

If  $k \in I_2$ , then

$$u_{k+1} \leq u_k - (1-\theta)\gamma_k\phi(u_k).$$

Two cases are possible:

(a)  $I_1$  is finite. Then  $k \in I_2$  and  $u_{k+1} \leq u_k$  for all sufficiently large  $k$ . Hence  $u_k \rightarrow \bar{u} \geq 0$ . Since  $\phi(u)$  is monotone,  $\phi(u_k) \geq \phi(\bar{u})$ ,  $u_{k+1} \leq u_k - (1-\theta)\gamma_k\phi(u_k)$ . Summing the inequalities and taking into account the assumption  $\sum \gamma_k = \infty$  we get  $\phi(\bar{u}) = 0$ , i.e.,  $\bar{u} = 0$ .  $\checkmark K$

(b)  $I_1$  is infinite. Then for  $k \in I_1$ ,  $\phi(u_k) \leq \varepsilon_k + \delta_k u_k$ ,  $\varepsilon_k = \beta_k / (\theta\gamma_k) \rightarrow 0$ ,  $\delta_k = \alpha_k / (\theta\gamma_k) \rightarrow 0$ . If  $\alpha_k \equiv 0$ , then  $\delta_k \equiv 0$ ,  $\phi(u_k) < \varepsilon_k \rightarrow 0$ , hence  $u_k \rightarrow 0$  for  $k \in I_1$ ,  $k \rightarrow \infty$ . If  $\alpha_k \neq 0$  but  $\phi(u)$  is convex, then the equation

$$\phi(u) = \varepsilon + \delta u$$

has a single solution  $u^*(\varepsilon, \delta) > 0$  for sufficiently small  $\varepsilon \geq 0$ ,  $\delta > 0$  and  $u^*(\varepsilon, \delta) \rightarrow 0$  for  $\varepsilon, \delta \rightarrow 0$ . If  $\phi(u) < \varepsilon + \delta u$ , then  $u < u^*(\varepsilon, \delta)$ . Hence  $u_k < u^*(\varepsilon_k, \delta_k)$  for  $k \in I_1$  and  $u_k \rightarrow 0$  for  $k \in I_1$ ,  $k \rightarrow \infty$ . Thus  $u_k \rightarrow 0$  for  $k \in I_1$ ,  $k \rightarrow \infty$  regardless  $\alpha_k \equiv 0$  or  $\alpha_k \neq 0$ . If  $k \in I_1$ ,  $k+1 \in I_2$ , then  $u_{k+1} \leq (1+\alpha_k)u_k + \beta_k \rightarrow 0$  for  $k \in I_1$ ,  $k \rightarrow \infty$ . Finally, if  $k \in I_1$ ,  $k+j \in I_2$ ,  $j = 1, \dots, s$ , then  $u_{k+j} \leq u_{k+j-1} \leq \dots \leq u_{k+1}$ . Hence  $u_k \rightarrow 0$ ,  $k \rightarrow \infty$ .

Thus  $u_k \rightarrow 0$  for both cases.  $\square$

## 2.2.2 Lemmas on Random Sequences

To investigate iterative methods which manifest random characteristics (the method of random search, problems with noise), one usually applies the

same technique based on Lyapunov functions. However, in this case, the Lyapunov function is a random variable: hence analogs of the preceding lemmas for random sequences are needed.

Recall the various forms of convergence of random variables. Let  $v^1, \dots, v^k, \dots$  be a sequence of  $n$ -dimensional random vectors. We shall not specify the probability space  $(\Omega, \mathcal{F}, P)$  on which these variables are defined (i.e., we do not write  $v^1(\omega), \dots, v^k(\omega), \omega \in \Omega$ ,  $\Omega$  being the space of elementary events,  $\mathcal{F}$  being the  $\sigma$ -algebra of measurable sets defined on it,  $P$  being the probability measure on  $\mathcal{F}$ ). We say that the sequence  $v^k$  converges to the random vector  $v$ :

a) *almost surely (with probability 1)*, if  $P(\lim_{k \rightarrow \infty} v^k = v) = 1$  (here and in the sequel,  $P(A)$  denotes the probability of the event  $A$ ), and we indicate this by  $v^k \rightarrow v$  a.s.;

b) *in probability*, if for each  $\varepsilon > 0$ ,  $\lim_{k \rightarrow \infty} P(\|v^k - v\| > \varepsilon) = 0$ ;  
 $v^k \xrightarrow{P} v$ ;

c) *in the mean square*, if  $\lim_{k \rightarrow \infty} E\|v^k - v\|^2 = 0$  (here and in the sequel,  
 $E_\alpha$  denotes the mathematical expectation of the random variable  $\alpha$ ).

The theory of semimartingales is the basic tool for studying convergence of random variables. A sequence of scalar random variables  $v_0, \dots, v_k, \dots$  is called a *supermartingale* if  $E(v_{k+1} | v_0, \dots, v_k) \leq v_k$ ,  $E v_0 < \infty$ , where  $E(v_{k+1} | v_0, \dots, v_k)$  is the conditional mathematical expectation of  $v_{k+1}$  for the given  $v_0, \dots, v_k$ . If the inequality is of the opposite sign, then it is called a *submartingale*, whereas for the equality the term *martingale* is used. A *semimartingale* is a generalization to the stochastic case of the notion of a monotonically decreasing sequence. The key result on convergence of numerical sequences (a monotone decreasing sequence that is bounded below has a limit) has the following form for random variables.

**LEMMA 7.** Let  $v_0, \dots, v_k, \dots$  be a supermartingale, where  $v_k \geq 0$  for all  $k$ . Then there is a random variable  $v \geq 0$  such that  $v_k \rightarrow v$  a.s.  $\square$

The well-known *Chebyshev inequality* (if  $v \geq 0$ ,  $\varepsilon > 0$ ,  $E v < \infty$ , then  $P(v \geq \varepsilon) \leq \varepsilon^{-1} E v$ ) can be strengthened.

**LEMMA 8** (Kolmogorov's inequality). Let  $v_0, \dots, v_k, \dots$  be a *semimartingale*,  $v_k \geq 0$ ,  $\varepsilon > 0$ . Then

$$P(\exists k: v_k \geq \varepsilon) \leq \varepsilon^{-1} E v_0 . \quad \square \quad (18)$$

Using these results, we get stochastic analogs of Lemmas 2 and 3.

**LEMMA 9** (Gladyshev). Let there be a sequence of random variables  $v_0, \dots, v_k \geq 0$ ,  $E v_0 < \infty$  and

$$E(v_{k+1} | v_0, \dots, v_k) \leq (1 + \alpha_k)v_k + \beta_k,$$

$$\sum_{k=0}^{\infty} \alpha_k < \infty, \quad \sum_{k=0}^{\infty} \beta_k < \infty, \quad \alpha_k \geq 0, \quad \beta_k \geq 0.$$

Then  $v_k \rightarrow v$  a.s., where  $v \geq 0$  is some random variable.

**PROOF.** Introduce

$$u_k = \prod_{i=k}^{\infty} (1 + \alpha_i)v_i + \sum_{i=k}^{\infty} \beta_i \times \prod_{j=i+1}^{\infty} (1 + \alpha_j).$$

Then  $u_k \geq 0$ ,  $E u_0 < \infty$  (since

$$\prod_{i=0}^{\infty} (1 + \alpha_i) < \infty, \quad \sum_{i=0}^{\infty} \beta_i < \infty, \quad E v_0 < \infty).$$

Here

$$\begin{aligned} E(u_{k+1} | u_0, \dots, u_k) \\ = \prod_{i=k+1}^{\infty} (1 + \alpha_i) E(v_{k+1} | v_0, \dots, v_k) + \sum_{i=k+1}^{\infty} \beta_i \prod_{j=i+1}^{\infty} (1 + \alpha_j) \\ \leq \prod_{i=k}^{\infty} (1 + \alpha_i)v_i + \sum_{i=k}^{\infty} \beta_i \prod_{j=i+1}^{\infty} (1 + \alpha_j) = u_k, \end{aligned}$$

+ super

i.e.,  $u_k$  is a semi-martingale and by Lemma 7,  $u_k \rightarrow u$  a.s.,  $u \geq 0$ . Hence also

$$v_k = \left[ u_k - \sum_{i=k}^{\infty} \beta_i \prod_{j=i+1}^{\infty} (1 + \alpha_j) \right] / \prod_{i=k}^{\infty} (1 + \alpha_i) \rightarrow v \text{ a.s. } \square$$

**LEMMA 10.** Let  $v_0, \dots, v_k$  be a sequence of random variables,  $v_k \geq 0$ ,  $E v_0 < \infty$  and let

$$E(v_{k+1} | v_0, \dots, v_k) \leq (1 + \alpha_k)v_k + \beta_k, \quad (20)$$

$$0 \leq \alpha_k \leq 1, \quad \beta_k \geq 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty, \quad \sum_{k=0}^{\infty} \beta_k < \infty,$$

$$\frac{\beta_k}{\alpha_k} \rightarrow 0. \quad (21)$$

Then  $v_k \rightarrow 0$  a.s.,  $\mathbf{E}v_k \rightarrow 0$ , and for every  $\varepsilon > 0$ ,  $k > 0$ ,

$$\mathbf{P}(v_j \leq \varepsilon \text{ for all } j \geq k) \geq 1 - \varepsilon^{-1} \left[ \mathbf{E}v_k + \sum_{i=k}^{\infty} \beta_i \right]. \quad (22)$$

**PROOF.** Taking the unconditional mathematical expectation on both sides of (20), we obtain

$$\mathbf{E}v_{k+1} \leq (1 - \alpha_k) \mathbf{E}v_k + \beta_k,$$

and by Lemma 3,  $\mathbf{E}v_k \rightarrow 0$ . On the other hand,  $u_k = v_k + \sum_{i=k}^{\infty} \beta_i$  is a ~~semi~~ martingale (cf. the proof of Lemma 9). Using Lemmas 8 and 9, we get the required result.  $\square$

In the preceding lemmas the quantities  $\alpha_k, \beta_k$  were deterministic. Consider the case when they are random (and perhaps dependent).

**LEMMA 11** (Robbins-Siegmund). Let  $v_k, u_k, \alpha_k, \beta_k$  be nonnegative random variables and let

$$\mathbf{E}(v_{k+1} | F_k) \leq (1 + \alpha_k)v_k - u_k + \beta_k \quad \text{a.s.}$$

$$\sum_{k=0}^{\infty} \alpha_k < \infty \quad \text{a.s.}, \quad \sum_{k=0}^{\infty} \beta_k < \infty \quad \text{a.s.},$$

where  $\mathbf{E}(v_{k+1} | F_k)$  denotes the conditional mathematical expectation for the given  $v_0, \dots, v_k, u_0, \dots, u_k, \alpha_0, \dots, \alpha_k, \beta_0, \dots, \beta_k$ . Then

$$v_k \rightarrow v \quad \text{a.s.}, \quad \sum_{k=0}^{\infty} u_k < \infty \quad \text{a.s.},$$

where  $v \geq 0$  is some random variable.  $\square$

### 2.2.3 The Main Theorems

Consider an iterative process of the form

$$x^{k+1} = x^k - \gamma_k s^k, \quad (23)$$

where  $k$  is the number of the iteration,  $x^k, s^k$  are vectors in  $\mathbf{R}^n$ ,  $\gamma_k \geq 0$  is a scalar factor characterizing the step size. We combine the deterministic and the stochastic cases and consider the general situation when  $x^k$  and  $s^k$  are random, with the deterministic cases included as a special

case. The basic assumptions concerning the process are the following:

- a) The process is Markov: the distribution of  $s^k$  depends only on  $x^k$  and  $k$ ,  $s^k = s^k(x^k)$ , the variables  $s^k, s^{k-1}, \dots$  are mutually independent.
- b) There is a scalar function (the *Lyapunov function*)  $V(x) \geq 0$ ,  $\inf V(x) = 0$ ,  $V(x)$  is differentiable and  $\nabla V(x)$  satisfies a Lipschitz condition with constant  $L$ .
- c) Process (23) is *pseudogradient* in relation to the  $V(x)$ :

$$(\nabla V(x^k), \mathbf{E}(s^k | x^k)) \geq 0, \quad (24)$$

i.e.,  $-s^k$  in the mean is a direction of decrease of  $V(x)$  to the point  $x^k$ .

- d) The following *growth condition* on  $s^k$  is satisfied:

$$\mathbf{E}(\|s^k\|^2 | x^k) \leq \sigma^2 + \tau(\nabla V(x^k), \mathbf{E}(s^k | x^k)). \quad (25)$$

The variable  $\sigma^2$  usually characterizes the level of additive noise. The case  $\sigma = 0$  is typical for deterministic problems.

- e) The initial approximation satisfies the condition

$$\mathbf{E}V(x^0) < \infty. \quad (26)$$

It goes without saying that this condition holds if  $x^0$  is a deterministic vector.

- f) The step size is such that

$$\gamma_k \geq 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty, \quad \overline{\lim}_{k \rightarrow \infty} \gamma_k < \frac{2}{L\tau}. \quad (27)$$

Let us state the basic convergence theorems. Under conditions a-f, it is generally impossible to assert that  $V(x^k) \rightarrow 0$  for process (23) in any probabilistic sense. For example, if  $s^k \equiv 0$ , then all the conditions hold, but  $x^k = x^0$ . However, certain convergence assertions are valid even under these minimal assumptions.

**THEOREM 1.** Let conditions a-f hold and let either  $\sigma^2 = 0$  or  $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$ . Then for any  $x^0$  in algorithm (23) one has

$$V(x^k) \rightarrow V \text{ a.s., } \underline{\lim}_{k \rightarrow \infty} (\nabla V(x^k), \mathbf{E}(s^k | x^k)) = 0. \quad (28)$$

**PROOF.** Condition b and formula (15) of Section 1.1 yield in this case

$$V(x^{k+1}) \leq V(x^k) - \gamma_k(\nabla V(x^k), s^k) + L\gamma_k^2 \|s^k\|^2/2.$$

Let us take the conditional mathematical expectation on both sides of this inequality and apply condition d:

$$\begin{aligned} & \mathbf{E}(V(x^{k+1}) | x^k) \\ & \leq V(x^k) - \gamma_k (\nabla V(x^k), \mathbf{E}(s^k | x^k)) + L\gamma_k^2 \mathbf{E}(\|s^k\|^2 | x^k)/2 \\ & \leq V(x^k) - \gamma_k \left(1 - \left(\frac{1}{2}\right)L\tau\gamma_k\right) (\nabla V(x^k), \mathbf{E}(s^k | x^k)) + L\gamma_k^2 \sigma^2/2. \end{aligned} \quad (29)$$

By conditions c and f we have

$$\mathbf{E}(V(x^{k+1}) | x^k) \leq V(x^k) + L\gamma_k^2 \sigma^2/2. \quad (30)$$

Applying Lemma 9, we obtain that  $V(x^k) \rightarrow V$  a.s. Let us pass to unconditional mathematical expectations in (29):

$$\begin{aligned} \mathbf{EV}(x^{k+1}) & \leq \mathbf{EV}(x^k) - \gamma_k \left(1 - \left(\frac{1}{2}\right)L\tau\gamma_k\right) u_k + L\gamma_k^2 \sigma^2/2, \\ u_k & = \mathbf{E}(\nabla V(x^k), \mathbf{E}(s^k | x^k)). \end{aligned}$$

For sufficiently large  $k$ , by condition f, we have

$$\mathbf{E}(\nabla V(x^{k+1})) \leq \mathbf{EV}(x^k) - \gamma_k \varepsilon u_k + L\gamma_k^2 \sigma^2/2.$$

Since  $\mathbf{EV}(x^0) < \infty$  (condition e) and  $\sigma^2 \sum_{k=0}^{\infty} \gamma_k^2 < \infty$ , then  $\sum_{k=0}^{\infty} \gamma_k u_k < \infty$ . But since  $\sum_{k=0}^{\infty} \gamma_k = \infty$ , this means that  $\lim_{k \rightarrow \infty} u_k = 0$ . It follows from the properties of convergence in the mean that if  $\mathbf{E}z^k \rightarrow \infty$  for the random variables  $z^k \geq 0$ , then we can find a subsequence  $z^{k_i} \rightarrow 0$  a.s. Hence

$$\lim_{k \rightarrow \infty} (\nabla V(x^k), \mathbf{E}(s^k | x^k)) = 0 \text{ a.s. } \square$$

Now let us replace condition c by condition c' for the strong pseudogradient:  $c')$

$$(\nabla V(x^k), \mathbf{E}(s^k | x^k)) \geq \ell V(x^k), \quad \ell > 0.$$

**THEOREM 2.** Let conditions a-f and c' hold and let either  $\sigma^2 = 0$  or  $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$ . Then for any  $x^0$  in algorithm (23) one has  $V(x^k) \rightarrow 0$  a.s.:

$$\mathbf{P}(V(x^i) \leq \varepsilon \ \forall i \geq k) \geq 1 - \varepsilon^{-1} \left( \mathbf{EV}(x^k) + \frac{1}{2} L \sigma^2 \sum_{i=k}^{\infty} \gamma_i^2 \right). \quad (31)$$

**PROOF.** From (29) and condition c' we have

$$\mathbf{E}(V(x^{k+1}) | x^k) \leq (1 - \ell\gamma_k(1 - (\frac{1}{2})L\tau\gamma_k))V(x^k) + L\gamma_k^2\sigma^2/2. \quad (32)$$

The required result follows from Lemma 10 and condition f.  $\square$

Next we turn to conditions for convergence in the mean.

**THEOREM 3.** Let conditions a-f, c' hold and let either  $\sigma^2 = 0$  or  $\gamma_k \rightarrow 0$ . Then in algorithm (23) one has

$$\mathbf{EV}(x^k) \rightarrow 0. \quad (33)$$

**PROOF.** Taking the unconditional mathematical expectation in (32) yields

$$\mathbf{EV}(x^{k+1}) \leq (1 - \ell\gamma_k(1 - (\frac{1}{2})L\tau\gamma_k))\mathbf{EV}(x^k) + L\gamma_k^2\sigma^2/2. \quad (34)$$

Since

$$1 - (\frac{1}{2})L\tau\gamma_k \geq \varepsilon > 0$$

for sufficiently large  $k$ , then

$$\mathbf{EV}(x^{k+1}) \leq (1 - \ell\varepsilon\gamma_k)\mathbf{EV}(x^k) + L\gamma_k^2\sigma^2/2.$$

By Lemma 3,  $\mathbf{EV}(x^k) \rightarrow 0$ .  $\square$

One can also derive from inequality (34) other results, including convergence-rate estimates. Here are examples.

**THEOREM 4.** Let conditions a-f, c' hold and let  $\gamma_k \equiv \gamma$ ,  $0 < \gamma < 2/(L\tau)$ . Then

$$\begin{aligned} \mathbf{EV}(x^k) &\leq \mathbf{EV}(x^0)q^k + \frac{L\gamma\sigma^2}{\ell(2-L\tau\gamma)}(1-q^k), \\ q &= 1 - \ell\gamma(1 - (\frac{1}{2})L\tau\gamma). \end{aligned} \quad (35)$$

This result follows from (34) and Lemma 1.  $\square$

Thus, if  $\sigma^2 > 0$ , then

$$\overline{\lim_{k \rightarrow \infty}} \mathbf{EV}(x^k) \leq L\tau\sigma^2/[\ell(2-L\tau\gamma)],$$

but if  $\sigma^2 = 0$ , then  $\mathbf{EV}(x^k)$  tends to 0 linearly.

**THEOREM 5.** Let conditions a-f, c' hold and let  $\sigma^2 > 0$  and  $\gamma_k = \gamma/k$ . Then

$$\mathbf{EV}(x^k) = \begin{cases} O(1/k) & \text{for } \ell\gamma > 1, \\ O(1/k^{\ell\gamma}) & \text{for } \ell\gamma < 1. \end{cases} \quad (36)$$

This result can easily be derived from (34) and Lemma 4.  $\square$

## Exercises

1. Derive Theorem 1 of Section 1.4 as a corollary of Theorem 1, taking  $V(x) = f(x) - f^*$ .
2. Use Theorem 4 to prove Theorems 2 and 3 of Section 1.4, taking  $V(x) = f(x) - f^*$ , or  $V(x) = \|x - x^*\|^2$ .

### 2.2.4 Possible Modifications

The convergence theorems we have studied are not the most exhaustive. They can be modified in various directions.

1. Conditions c, c' and d can be generalized as follows:

$$(\nabla V(x^k), \mathbf{E}(s^k | x^k)) \geq \ell_k V(x^k) - \beta_k, \quad (37)$$

$$\mathbf{E}(\|s_k\|^2 | x^k) \leq \sigma_k^2 + \tau_k (\nabla V(x^k), \mathbf{E}(s^k | x^k)) + \mu_k V(x^k). \quad (38)$$

Under certain conditions imposed on  $\ell_k$ ,  $\beta_k$ ,  $\sigma_k$ ,  $\tau_k$  and  $\mu_k$ , using lemmas of this subsection, one can prove analogs of Theorems 1-3. We shall encounter conditions such as (37) and (38) while studying finite-difference variants of the gradient method, or regularization methods, among others.

2. All the results obtained so far have been global—we assumed that the conditions on  $V(x)$ ,  $s^k(x)$ , and so on, held for all  $x$  and the initial approximation could be arbitrary. However such assumptions often hold only locally, in a neighborhood of the solution. It is natural in this case that the convergence assertions be of a local nature, which is illustrated by Theorem 4 of Section 1.4 and Theorem 1 of Section 1.5 on local convergence of the gradient method and of Newton's method, respectively. Random noise complicates the situation—there is a nonzero probability of exit from the region in which the assumptions are satisfied. Hence the assertion on local convergence can hold only with some probability  $1 - \delta$ ,  $\delta > 0$ . We give now the corresponding analog of Theorem 2. Let

$$Q = \{x: V(x) \leq \varepsilon\},$$

where  $\varepsilon > 0$  is an integer.

**THEOREM 6.** Let conditions a-f, c' hold for all  $x, x^k \in Q$ . Then for method (23):

a) if  $x^0$  is deterministic, and  $x^0 \in Q$ ,  $\sigma^2 = 0$  and  $s^k$  is deterministic, then  $V(x^k) \rightarrow 0$ ;

b) if  $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$ , then

$$\mathbf{P}(x^k \in Q \ \forall k) \geq 1 - \delta, \quad \mathbf{P}(V(x^k) \rightarrow 0) \geq 1 - \delta, \quad (39)$$

$$\delta = \varepsilon^{-1} \mathbf{EV}(x^0) + \frac{1}{2} L\sigma^2 \varepsilon^{-1} \sum_{k=0}^{\infty} \gamma_k^2. \quad \square$$

One can consider *continuous-time analogs of iterative methods*—processes described by ordinary differential equations

$$dx/dt = s(x, t), \quad x(0) = x^0. \quad (40)$$

The same technique based on the Lyapunov function can be used for these analogs. Formulation of many convergence theorems becomes simpler and acquires a more intuitive meaning. Historically, the method of Lyapunov functions was originally developed for such problems. However, we do not give the corresponding results, nor examine continuous-time methods. The point is that we use digital computers to solve computational problems, and in any implementation of the process (4) on a computer one has to go over to a discrete-time approximation. Still, one needs to bear in mind that a transition to the “limiting” form of a discrete trajectory can be appropriate, from a methodological point of view, to simplify the formulations and to “predict” different methods. To prove the convergence, such an approach has been used systematically in Belen'kij, et al. [2.1].

Finally, one often examines an iterative process of the form

$$x^{k+1} = T(x^k), \quad T: \mathbf{R}^n \rightarrow \mathbf{R}^n, \quad (41)$$

rather than the form (23). The existence of a function  $V(x)$  with the property

$$V(T(x)) < V(x), \quad x \neq T(x), \quad (42)$$

is postulated and  $V(x)$  and  $T(x)$  are required to be neither differentiable nor smooth. It suffices to assume, for instance, that the function  $\phi(x) = V(T(x))$  is lower semicontinuous and the set  $\{x: V(x) \leq V(x^0)\}$  is bounded. Under these conditions, it is possible to prove that sequence (41) has limit points, each of which is a fixed point of  $T(x)$ . Schemes of this kind have been suggested and investigated in [0.6], [0.13], [1.6], [2.9]. The scheme developed by E.A. Nurminskij in [2.9] is promising along these

lines: the  $V(x)$  are not required to be monotonically decreasing at each step and the scheme is applicable to the stochastic case as well. Unfortunately such approaches provide no information relating to the rate of convergence of the process.

## 2.3 OTHER SCHEMES

One should not think that the first and second Lyapunov method exhaust the whole variety of schemes for investigating convergence of iterative procedures. These schemes are sometimes based on different considerations. Let us briefly describe some of them.

### 2.3.1 The Contraction Mapping Principle

Let  $g: \mathbf{R}^n \rightarrow \mathbf{R}^n$  be a mapping. It is called a *contraction mapping* if

$$\|g(x) - g(y)\| \leq q\|x - y\|, \quad q < 1, \quad (1)$$

for all  $x, y \in \mathbf{R}^n$ , i.e., if it satisfies a Lipschitz condition with a constant smaller than 1. Consider the iterative process

$$x^{k+1} = g(x^k). \quad (2)$$

**THEOREM 1** (the contraction mapping principle). If  $g$  is a contraction mapping, then it has a unique fixed point  $x^*$  to which process (2) converges for any  $x^0$  with the rate of geometric progression:

$$\|x^k - x^*\| \leq q^k(1 - q)^{-1} \|g(x^0) - x^0\|. \quad (3)$$

**PROOF.**

$$\begin{aligned} \|x^{k+1} - x^k\| &= \|g(x^k) - g(x^{k-1})\| \leq q\|x^k - x^{k-1}\|, \\ \|x^{k+1} - x^k\| &\leq q^k\|x^1 - x^0\|, \\ \|x^{k+s} - x^k\| &\leq \sum_{i=k}^{k+s-1} \|x^{i+1} - x^i\| \\ &\leq (q^{k+s-1} + q^{k+s-2} + \dots + q^k)\|x^1 - x^0\| \\ &\leq \frac{q^k}{1 - q} \|x^1 - x^0\|. \end{aligned} \quad (4)$$

Hence,  $\|x^{k+s} - x^k\| \rightarrow 0$  as  $k \rightarrow \infty$  for any  $s$ , i.e.,  $x^k$  is a Cauchy sequence in  $\mathbf{R}^n$ . Since  $\mathbf{R}^n$  is complete,  $x^k$  has a limit  $x^*$ . Since  $g(x)$  is continuous

by (1), it follows from  $x^k \rightarrow x^*$  that  $g(x^k) \rightarrow g(x^*)$ , but  $g(x^k) = x^{k+1} \rightarrow x^*$ . Hence  $x^* = g(x^*)$ . Passing to the limit in (4) as  $s \rightarrow \infty$ , we get  $\|x^* - x^k\| \leq (q^k/(1-q)) \|x^1 - x^0\|$ . The uniqueness of a fixed point follows from (1) immediately.  $\square$

The contraction mapping principle is convenient because it asserts the convergence of the iterative process, as well as guarantees the existence of a fixed point. That is why it has usually been applied in mathematics for deriving various existence theorems.

This principle has many different realizations and modifications. Yet it cannot essentially be extended, as Exercises 1-3 below demonstrate.

We also note that an attempt to apply the contraction mapping principle to the problems considered in Section 2.1 does not pay off. Indeed, we proved therein that if the spectral radius  $\rho(A)$  of the matrix  $A$  is less than 1, then the iterations  $x^{k+1} = Ax^k$  converge. However, under these conditions, the linear mapping  $g(x) = Ax$  is, generally, not a contraction, since one does not necessarily have  $\|A\| < 1$ ; see Section 2.1.

## Exercises

1. Construct an example of a mapping  $g(x)$  having the property:  $\|g(x) - g(y)\| < \|x - y\|$  for any  $x \neq y$ , but not having a fixed point.
2. Construct an example of a nonexpanding mapping:  $\|g(x) - g(y)\| < \|x - y\|$  with a fixed point, for which the iterations  $x^{k+1} = g(x^k)$  do not converge.
3. Construct an example of contraction mappings  $g_k$  with the same contraction constant  $q < 1$ , for which the iterations  $x^{k+1} = g_k(x^k)$  do not converge.

### 2.3.2 The Implicit Function Theorem

A convenient tool for investigating iterative methods not explicit with respect to  $x^{k+1}$  is the well-known implicit function theorem from Analysis. Let  $F(x, y)$  be a mapping from  $\mathbf{R}^n \times \mathbf{R}^n$  to  $\mathbf{R}^n$ . We denote by  $F'_x(x, y), F'_y(x, y)$  the derivatives of  $F$  with respect to the corresponding variables.

**THEOREM 2** (implicit function theorem). Let  $F(x^*, y^*) = 0$ , let  $F(x, y)$  be continuous with respect to  $\{x, y\}$  in a neighborhood of  $x^*, y^*$ , differentiable in  $x$  in a neighborhood of  $x^*, y^*$ , let  $F'_x(x, y)$  be continuous at  $x^*, y^*$ , and let the matrix  $F'_x(x^*, y^*)$  be nonsingular. Then there exists a unique function  $x = \phi(y)$  continuous in a neighborhood of  $y^*$ , such that  $x^* = \phi(y^*)$ ,  $F(\phi(y), y) = 0$ . Moreover, if  $F'_y(x^*, y^*)$  exists, then  $\phi(y)$  is differentiable at  $y^*$  and

$$\phi'(y^*) = -[F'_x(x^*, y^*)]^{-1} F'_y(x^*, y^*) . \quad \square \quad (5)$$

In other words, the equation  $F(x, y) = 0$  can be solved for  $x$  in a neighborhood of  $y^*$ . We apply this result first to investigate the existence and uniqueness of solutions.

**THEOREM 3.** Let the equation  $g(x) = 0$ ,  $g: \mathbf{R}^n \rightarrow \mathbf{R}^n$ , have a solution  $x^*$ , where  $g(x)$  is differentiable in a neighborhood of  $x^*$ ,  $g'(x)$  is continuous at  $x^*$  and the matrix  $g'(x^*)$  is nonsingular. Then the equation

$$g(x) = y \quad (6)$$

has a solution  $x(y)$  for sufficiently small  $y$ , and

$$x(y) = x^* - g'(x^*)^{-1}y + o(y). \quad \square \quad (7)$$

These results allow us to investigate iterative processes in which the new approximation  $x^{k+1}$  is an implicit expression—for example, it may be a solution of some auxiliary problem of unconstrained minimization. This is what we observe in the regularization method and in many methods of unconstrained minimization, the penalty-function method, among others.

### 2.3.3 The Role of General Schemes for Investigating Convergence

General theorems of the type described in this chapter take upon themselves the standard, routine part of proving the convergence; they thereby simplify the proof of algorithms. However, one should not exaggerate their significance and assume that they make convergence analysis elementary. First, in many cases, verification of the conditions is an independent and nontrivial problem. Secondly, for simple problems a direct—“frontal”—proof is in no way more complex than specializing general theorems. We saw examples in Chapter 1. Of course, one could prove those results, using the arguments of this chapter, but they are not as obvious and instructional as the direct proofs. Finally, in some problems it is advantageous to employ special techniques exploiting particular features of the problem.

An analysis of convergence still remains a challenging and creative procedure which calls for artistry as well as common sense. Attempts to procrusteanize this procedure into a well-cut unified scheme—as is characteristic of certain monographs—have not been fruitful.

## CHAPTER 3

### MINIMIZATION METHODS

In Chapter 1 we considered two minimization algorithms that are conceptually the simplest: the gradient method and Newton's method. There are many other methods of unconstrained minimization of differentiable functions. We shall describe the most interesting ones—either theoretically or computationally. Throughout this chapter we shall specialize to the problem

$$\min f(x), \quad x \in \mathbf{R}^n,$$

where  $f(x)$  is a differentiable function.

#### 3.1 MODIFICATIONS OF THE GRADIENT METHOD AND OF NEWTON'S METHOD

##### 3.1.1 Advantages and Drawbacks of the Earlier Methods

In Chapter 1 we discussed in detail the gradient method

$$x^{k+1} = x^k - \gamma \nabla f(x^k) \quad (1)$$

and Newton's method

$$x^{k+1} = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k). \quad (2)$$

In Table 1 we list the advantages versus drawbacks of each method (the terminology was explained in Chapter 1). As is seen from Table, the posi-

tive and negative features of each method are complementary. Of course it would be ideal to develop a new method combining the best features, eschewing the disadvantages. Although such an ideal solution does not exist, we shall describe now some possible steps toward it.

It turns out that some of the drawbacks—the need to choose  $\gamma$  for the gradient method, the local nature of Newton's method—can be eliminated by a simple modification of the methods.

TABLE 1

Method	Advantages	Drawbacks
Gradient	Global convergence. Relaxed conditions on $f(x)$ . Computational simplicity	Slow convergence. Necessary choice of $\gamma$
Newton's	Rapid convergence	Local convergence. Rigid conditions on $f(x)$ . Large volume of computation

### 3.1.2 Modifications of the Gradient Method

Let us consider the general gradient method

$$x^{k+1} = x^k - \gamma_k \nabla f(x^k) \quad (3)$$

for various ways of choosing the step size  $\gamma_k$ . At first, it seems possible to improve significantly the efficiency of the gradient method by going to a minimum in the antigradient direction:

$$\gamma_k = \underset{\gamma \geq 0}{\operatorname{argmin}} \phi_k(\gamma), \quad \phi_k(\gamma) = f(x^k - \gamma \nabla f(x^k)). \quad (4)$$

We have thus obtained the so-called *steepest descent method*.

**THEOREM 1.** Let  $f(x)$  be a continuously differentiable function and let  $\{x: f(x) \leq f(x^0)\}$  be bounded. Then in method (3), (4),  $\nabla f(x^k) \rightarrow 0$  and the sequence  $x^k$  has limit points each of which is stationary, i.e., we can find the subsequence  $x^{k_i} \rightarrow x^*$ , and  $\nabla f(x^*) = 0$ .

This result is not hard to prove, using the technique given in Chapter 2. In contrast to Theorem 1 of Section 1.4, the Lipschitz condition on the gradient can be replaced by a weaker condition of gradient continuity. This is natural to do, since the choice of the step size (4) is less restrained than that of  $\gamma_k \equiv \gamma$ . Method (3), (4) converges in the situations described in Section 1.4 to demonstrate the divergence of the gradient method with constant step if the Lipschitz condition is not satisfied.  $\square$

Let us elucidate on the rate of convergence of the method. We consider the quadratic function

$$f(x) = (Ax, x)/2 - (b, x), \quad A > 0. \quad (5)$$

In (4), the  $\gamma_k$  can be written explicitly:

$$\gamma_k = \frac{\|\nabla f(x^k)\|^2}{(A\nabla f(x^k), \nabla f(x^k))}. \quad (6)$$

Method (3), (6) has the advantage over method (1) in that it does not contain the parameter  $\gamma$  subject to choice.

**THEOREM 2.** For method (3), (6) for the function (5) one has the estimate

$$f(x^k) - f(x^*) \leq (f(x^0) - f(x^*)) \left[ \frac{L-\ell}{L+\ell} \right]^{2k}, \quad (7)$$

where  $\ell$  and  $L$  are respectively the smallest and the largest eigenvalues of the matrix  $A$ ,  $x^* = A^{-1}b$  is a minimum point of  $f(x)$ .

**PROOF.** Using  $\phi_k(\gamma)$  and  $\gamma_k$ , we have

$$\begin{aligned} f(x^{k+1}) &= f(x^k) - \gamma_k (\nabla f(x^k), \nabla f(x^k)) + \gamma_k^2 (A \nabla f(x^k), \nabla f(x^k))/2 \\ &= f(x^k) - \frac{1}{2} \frac{\|\nabla f(x^k)\|^4}{(A \nabla f(x^k), \nabla f(x^k))}. \end{aligned}$$

Since

$$2(f(x^k) - f(x^*)) = (A(x^k - x^*), x^k - x^*) = (A^{-1} \nabla f(x^k), \nabla f(x^k)),$$

then

$$\frac{f(x^{k+1}) - f(x^*)}{f(x^k) - f(x^*)} = 1 - \frac{\|\nabla f(x^k)\|^4}{(A^{-1} \nabla f(x^k), \nabla f(x^k))(A \nabla f(x^k), \nabla f(x^k))}.$$

Using Kantorovich's inequality

$$(Ax, x)(A^{-1}x, x) \leq (4L\ell)^{-1}(L + \ell)^2 \|x\|^4 \quad \forall x \in \mathbf{R}^n, \quad (8)$$

we obtain

$$\frac{f(x^{k+1}) - f(x^*)}{f(x^k) - f(x^*)} \leq \left(\frac{L-\ell}{L+\ell}\right)^2$$

yielding the required estimate (7).  $\square$

Since

$$\nabla(f(x) - f(x^*)) = (A(x-x^*), x-x^*) \geq \ell\|x-x^*\|^2,$$

it follows from (7) that

$$\|x^k - x^*\| \leq \sqrt{2\ell^{-1}(f(x^0) - f(x^*))} q^k, \quad q = (L - \ell)/(L + \ell). \quad (9)$$

Estimate (7) is exact, since it is not hard to construct a two-dimensional example for which an inequality in (7) becomes equality. Comparing (7) and (9) with Theorem 3 of Section 1.4 leads to a somewhat unexpected conclusion: the steepest descent method for a quadratic function generally converges no faster than the simple gradient method (1) for the appropriate choice of  $\gamma$ . The same conclusion is also valid for the general nonquadratic case. Thus, in the gradient method we cannot improve the rate of convergence through a more complete one-dimensional minimization (i.e., by choosing the step size according to (4)).

We should not infer, however, that this can never be done. For example, if in order to minimize the quadratic function (5) we apply the gradient method (3) with  $\gamma_k = 1/\lambda_{k+1}$ ,  $k = 0, \dots, n-1$ , where  $\lambda_i$  are the eigenvalues of  $A$ , the method will be finite, i.e.,  $x^n = x^*$ . (Verify!) Of course this result is hardly of practical interest, because the eigenvalues of  $A$  are usually unknown and finding them is a more difficult problem than solving the system  $Ax = b$ .

Let us look at another way of choosing  $\gamma_k$ . The simplest choice  $\gamma_k \equiv \gamma$ ,  $0 < \gamma < 2/L$  (Theorem 1 of Section 1.4) is nonconstructive since the constant  $L$  is usually unknown. The following procedure for choosing  $\gamma$  can be advantageous. Let  $0 < \varepsilon < 1$ ,  $0 < \alpha < 1$  and let some  $\gamma$  be given. We compute  $f(x^k - \gamma \nabla f(x))$  and verify the inequality

$$f(x^k - \gamma \nabla f(x^k)) \leq f(x^k) - \varepsilon \gamma \|\nabla f(x^k)\|^2 \quad (10)$$

in each iteration. If it is satisfied, then  $x^{k+1} = x^k - \gamma \nabla f(x^k)$ ; if not, then  $\gamma$  is replaced by  $\underline{\gamma_\alpha}$  and the check is repeated.

One can show that under the conditions of Theorems 1 and 2 of Section 1.4 this procedure requires a finite number of reductions of  $\gamma$  in each iteration and the assertions of these theorems remain in force. Thus, it is not hard to make the rule for choosing the step size to be constructive. Yet the main drawback of the gradient method, that is, its poor convergence for ill-posed problems, cannot be removed by simple means.

### 3.1.3 Modifications of Newton's Method

One can make Newton's method globally convergent in various ways. One of them involves regulating the step size:

$$x^{k+1} = x^k - \gamma_k [\nabla^2 f(x^k)]^{-1} \nabla f(x^k). \quad (11)$$

It is often called the *damped Newton's method*. The parameter  $\gamma_k$  can be selected in different ways, for example,

$$\gamma_k = \underset{\gamma \geq 0}{\operatorname{argmin}} f(x^k - \gamma [\nabla^2 f(x^k)]^{-1} \nabla f(x^k)), \quad (12)$$

or  $\gamma$  is reduced (multiplied by  $0 < \alpha < 1$ ) beginning with  $\gamma = 1$  until the condition

$$f(x^{k+1}) \leq f(x^k) - \gamma q ([\nabla^2 f(x^k)]^{-1} \nabla f(x^k), \sqrt{f(x^k)}), \quad 0 < q < 1, \quad (13) \quad \checkmark$$

or the condition

$$\|\nabla f(x^{k+1})\|^2 \leq (1 - \gamma q) \|\nabla f(x^k)\|^2, \quad 0 < q < 1, \quad (14)$$

is satisfied.

For smooth strongly convex functions the damped Newton's method converges globally (Exercise 1). In the initial iterations the rate of convergence can be only linear, but as soon as the method arrives in a neighborhood of  $x^*$ , in which the conditions of Theorem 1 of Section 1.5 are satisfied, the rate becomes quadratic (Exercise 2).

Another modification—the so-called Levenberg-Marquardt method—is also possible, in which the actual direction differs from the one given by Newton's method. We proceed as we did in justifying the gradient method (see (3) in Section 1.4), viz. we add to the approximating function a quadratic penalty for deviating from the point  $x^k$ , i.e., we seek the  $x^{k+1}$  from the minimum condition

$$f_k(x) + (\alpha_k/2) \|x - x^k\|^2,$$

$$f_k(x) = f(x^k) + (\nabla f(x^k), x - x^k) + (\nabla^2 f(x^k)(x - x^k), x - x^k)/2, \quad (15)$$

$$x^{k+1} = x^k - (\nabla^2 f(x^k) + \alpha_k I)^{-1} \nabla f(x^k). \quad (16)$$

For  $\alpha_k = 0$  the method becomes Newton's; as  $\alpha_k \rightarrow \infty$  the direction tends to the antigradient. Thus (16) is a compromise between these two methods. By appropriately choosing  $\alpha_k$  one can make the method converge globally (Exercise 3).

Method (16) has the advantage over method (11) that it, as well as the gradient method, is not for convex functions only (see Exercise 3), whereas method (11) requires that the matrix  $\nabla^2 f(x^k)$  be positive definite (Exercise 4).

There are special modifications of Newton's method in which the matrix  $\nabla^2 f(x^k)$  is replaced by a positive definite matrix (if  $\nabla^2 f(x^k)$  is not).

However, in all of the modifications of Newton's method mentioned above, each iteration, as well as in the basic Newton's method, involves a large amount of computations (the computation of  $\nabla^2 f(x)$ , solving systems of linear equations), and the convergence rate is far from the minimum and, in general, low.

The attempts to "fix" the gradient method as well as Newton's method remove some of the drawbacks, but not the major ones, viz. poor convergence of the gradient method and intricate and laborious implementation of Newton's method.

## Exercises

- Let  $f(x)$  be a twice-differentiable strongly convex function,  $\|\nabla^2 f(x)\| \leq L$ . Then in procedures (13), (14) the number of reductions of  $\gamma$  in each iteration is finite, and method (11), with any rule (12)-(14) for choosing  $\gamma_k$  and for any  $x^0$ , converges to the minimum point  $x^*$  with a linear rate. Prove this, using the theorems of Section 2.2 and taking  $V(x) = f(x) - f(x^*)$  or  $V(x) = \|\nabla f(x)\|^2$ .
- Show that under the conditions of Theorem 1 of Section 1.5 in methods (13) and (14) one will have  $\gamma_k = 1$  in a sufficiently small neighborhood of  $x^*$ .
- Let  $f(x)$  be a twice-differentiable function, let  $\|\nabla^2 f(x)\| \leq L$ , let the set  $\{x: f(x) \leq f(x^0)\}$  be bounded, and let the point  $x^*$ , at which  $\nabla f(x^*) = 0$ , be unique. Show that one can find  $\underline{\gamma}$  and  $\bar{\gamma}$  such that for  $\underline{\gamma} \leq \alpha_k \leq \bar{\gamma}$ , in method (16) one will have  $x^k \rightarrow x^*$ . Use the theorems of Section 2.2 and take  $V(x) = f(x) - f(x^*)$ .
- Give examples showing that if the matrix  $\nabla^2 f(x^k)$  is not positive definite, method (11) may lose its meaning ( $[\nabla^2 f(x^k)]^{-1}$  does not exist) and that in method (11), (12) the  $\gamma_k$  might be equal to zero at a point at which  $\nabla f(x^k) \neq 0$ .

## 3.2 MULTISTEP METHODS

In the gradient method, at each step the information obtained in the preceding iterations is not used at all. It is natural to try to take into account the “prehistory” of the process in order to improve the convergence. Methods in which the new approximation depends on the  $s$  preceding ones:

$$x^{k+1} = \phi_k(x^k, \dots, x^{k-s+1}), \quad (1)$$

are called  $s$ -step methods. The gradient method and Newton's method are one-step methods. Next we shall consider multistep ( $s > 1$ ) methods.

### 3.2.1 The Heavy Ball Method

One of the simplest multistep methods is the two-step heavy-ball method

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}), \quad (2)$$

where  $\alpha > 0$ ,  $\beta \geq 0$  are parameters. Clearly, for  $\beta = 0$ , method (2) turns into the gradient method. The method owes its name to the following physical analogy. The motion of a body (“the heavy ball”) in a potential field under the force of friction (or viscosity) is described by a second-order differential equation

$$\mu \frac{d^2 x(t)}{dt^2} = -\nabla f(x(t)) - p \frac{dx(t)}{dt}. \quad (3)$$

Clearly, because of energy loss caused by friction, the body ultimately reaches a minimum point of the potential  $f(x)$ . Thus, the heavy ball “solves” the corresponding minimization problem. If we consider the difference analog of equation (3), we arrive at the iterative method (2).

The inertia (the term  $\beta(x^k - x^{k-1})$ ) introduced into the iterative process may increase the convergence. This is seen, for instance, from Figure 6: instead of the zigzag motion in the gradient method, the heavy-ball method has a smoother trajectory along the “bottom of the gully.” These heuristic considerations are strengthened by the following theorem.

**THEOREM 1.** Let  $x^*$  be a nonsingular minimum point of  $f(x)$ ,  $x \in \mathbb{R}^n$ . Then for

$$0 \leq \beta < 1, \quad 0 < \alpha < 2(1+\beta)/L, \quad \ell I \leq \nabla^2 f(x^*) \leq L I \quad (4)$$

we can find an  $\varepsilon > 0$  such that for any  $x^0, x^1$ ,  $\|x^0 - x^*\| \leq \varepsilon$ ,  $\|x^1 - x^*\| \leq \varepsilon$ , method (2) converges to  $x^*$  with the rate of geometric progression:

$$\|x^k - x^*\| \leq c(\delta)(q + \delta)^k, \quad 0 \leq q < 1, \quad 0 < \delta < 1-q. \quad (5)$$

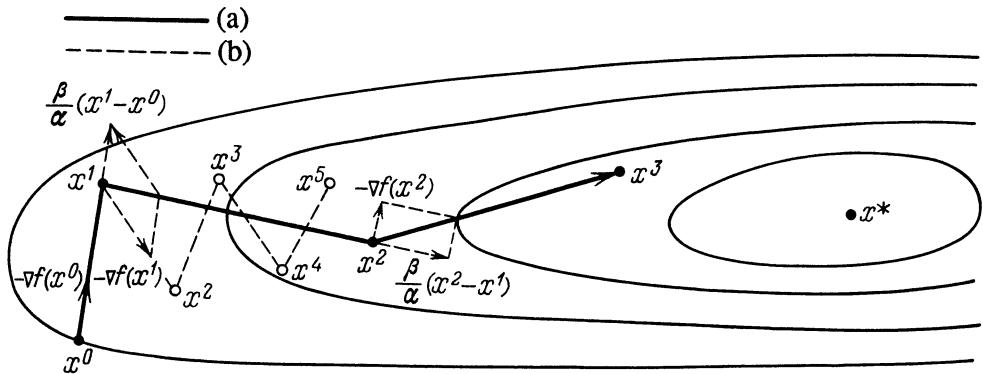


Fig. 6 (a) The heavy-ball method; (b) the gradient method.

The quantity  $q$  is minimal and equal to

$$q^* = \frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}} \quad \text{for } \alpha^* = \frac{4}{(\sqrt{L} + \sqrt{\ell})^2}, \quad \beta^* = \left( \frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}} \right)^2. \quad (6)$$

Sketch of a proof. In this case we cannot apply the procedures described in Chapter 2 for investigating the convergence since they are designed for one-step processes. We can, however, increase the dimension of the space which allows us to reduce the multistep process to a one-step process (see (15) in Section 2.1). Introduce the  $2n$ -dimensional vector  $z^k = \{x^k - x^*, x^{k-1} - x^*\}$ . Then the iterative process (2) can be written in the form

$$z^{k+1} = Az^k + o(z^k), \quad (7)$$

where the  $2n \times 2n$ -square matrix  $A$  has the form

$$A = \begin{bmatrix} (1+\beta)I - \alpha B & -\beta I \\ I & 0 \end{bmatrix}, \quad B = \nabla^2 f(x^*). \quad (8)$$

√ Let  $\ell = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n = L$  be the eigenvalues of the matrix  $B$ . Then

the eigenvalues  $\rho_j$ ,  $j = 1, \dots, 2n$ , of the matrix  $A$  coincide with the eigenvalues of  $2 \times 2$ -matrix of the form

$$\begin{bmatrix} 1 + \beta - \alpha \lambda_i & -\beta \\ 1 & 0 \end{bmatrix}.$$

Therefore, they are roots of the equations

$$\rho^2 - \rho(1 + \beta - \alpha \lambda_i) + \beta = 0, \quad i = 1, \dots, n. \quad (9)$$

One can show that if

$$0 < \ell \leq \lambda_i \leq L, \quad 0 \leq \beta < 1, \quad 0 < \alpha < 2(1+\beta)/L,$$

then  $|\rho| < 1$ , where  $\rho$  is any root of equation (9).

Now we can use Theorem 1 of Section 2.1 on local convergence of iterative processes of the form (7), which will allow us to obtain an estimate of (5). Calculating  $\min_{\alpha, \beta} \max_{1 \leq j \leq 2n} |\rho_j|$ , yields the optimal values  $\alpha^*$ ,  $\beta^*$  and the corresponding  $q^*$  given in the theorem.

Let us compare now the rate of convergence in the one-step and two-step methods for an optimal choice of parameters. In both cases we have the geometric rate of convergence, but the progression ratio for the one-step method is equal to

$$q_1 = (L - \ell)/(L + \ell), \quad (10)$$

whereas for the two-step method it is equal to

$$q_2 = (\sqrt{L} - \sqrt{\ell})/(\sqrt{L} + \sqrt{\ell}). \quad (11)$$

For large values of the condition number  $\mu = L/\ell$

$$q_1 \approx 1 - 2/\mu, \quad q_2 \approx 1 - 2/\sqrt{\mu}. \quad (12)$$

Hence, to be  $e = 2.7, \dots$  times closer to a solution, the one-step method takes roughly  $\mu/2$  iterations, and the two-step method roughly  $\sqrt{\mu}/2$  iterations. In other words, for ill-posed problems the heavy-ball method yields a roughly  $\sqrt{\mu}$ -fold payoff vs. the gradient method. For large  $\mu$  this difference is quite large. From the computational viewpoint, method (2) is only slightly more complex than the one-step method. Of course, a choice of optimal values for  $\alpha$  and  $\beta$  in (2) is not simple: we cannot directly use formulas (6), since the bounds of the spectrum of  $\nabla^2 f(x^*)$  (the numbers  $\ell$  and  $L$ ) are usually unknown.  $\square$

**Exercise**

1. Prove the global convergence for method (2) for a quadratic  $f(x)$ .

**3.2.2 The Conjugate Gradient Method**

Let us examine another variant of the two-step method—the conjugate gradient method, in which the parameters are found through solving the two-dimensional optimization problem:

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k(x^k - x^{k-1}), \quad (13)$$

$$\{\alpha_k, \beta_k\} = \underset{\{\alpha, \beta\}}{\operatorname{argmin}} f(x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1})). \quad (14)$$

For a quadratic function

$$f(x) = (Ax, x)/2 - (b, x), \quad A > 0, \quad (15)$$

this problem can be solved explicitly:

$$\begin{aligned} \alpha_k &= \frac{\|r^k\|^2(Ap^k, p^k) - (r^k, p^k)(Ar^k, p^k)}{(Ar^k, r^k)(Ap^k, p^k) - (Ar^k, p^k)^2}, & r^k &= \nabla f(x^k) = Ax^k - b, \\ \beta_k &= \frac{\|r^k\|^2(Ar^k, p^k) - (r^k, p^k)(Ar^k, r^k)}{(Ar^k, r^k)(Ap^k, p^k) - (Ar^k, p^k)^2}, & p^k &= x^k - x^{k-1}. \end{aligned} \quad (16)$$

One might expect that the relationship between methods (13), (14) and (2) is similar to that between methods (3), (4) and (1) of Section 3.1: the steepest descent method does not yield a higher convergence rate than the gradient method with constant optimal  $\gamma$ ; it is even less possible that a two-step variant of the steepest descent method (13), (14) may provide a substantially faster convergence than the heavy-ball method (2). This is not the case, however: in the quadratic case, method (13), (14) (for a special choice of  $p'$ ) is finite, i.e., it yields an exact minimum of the function (15) in a finite number of iterations.

Let the initial approximation  $x^0$  be arbitrary, and let  $x^1$  be obtained via the steepest descent method:

—

$$x^1 = x^0 - \frac{\|r^0\|^2}{(Ar^0, r^0)} r^0, \quad r^0 = \nabla f(x^0) = Ax^0 - b. \quad (17)$$

**LEMMA 1.** The gradients  $r^0, r^1, \dots$  in each method, (13), (16), (17), are pairwise orthogonal:

$$(r^i, r^k) = 0, \quad i < k. \quad (18)$$

**PROOF.** We use induction on  $k$ . Let  $(r^i, r^k) = 0$  for  $0 \leq i < k$ ,  $k \geq 2$ , and  $r^i \neq 0$ ,  $i = 0, \dots, k$ . The orthogonality of  $r^0, r^1, r^2$  follows directly from the definition of the method. Multiplying (13) on the left by  $A$  yields

$$r^{k+1} = r^k - \alpha_k A r^k + \beta_k (r^k - r^{k-1}).$$

It follows from  $r^i \neq 0$  for  $i \leq k$  that  $\alpha_k \neq 0$ . Hence  $A r^k$  is a linear combination of  $r^{k+1}, r^k$ , and  $r^{k-1}$ , and similarly  $A r^i$ ,  $i < k$ , is a linear combination of  $r^{i+1}, r^i, r^{i-1}$ , and by induction,  $(A r^i, r^j) = 0$ ,  $|i-j| > 1$ ,  $i < k$ ,  $j \leq k$ . Therefore

$$(r^{k+1}, r^i) = (r^k - \alpha_k A r^k + \beta_k (r^k - r^{k-1}), r^i) = 0 \quad \text{for } i = 0, \dots, k-2.$$

It follows directly from formulas (13), (16) that

$$(r^{k+1}, r^k) = 0, \quad (r^{k+1}, p^k) = 0.$$

Finally, from (13), replacing  $k$  by  $k-1$ , we have  $p^k = -\alpha_{k-1} r^{k-1} + \beta_{k-1} p^{k-1}$ . Applying this relation successively, we obtain that  $p^k$  is a linear combination of  $r^0, r^1, \dots, r^{k-1}$ , and  $r^{k-1}$  has the coefficient  $-\alpha_{k-1} \neq 0$ . Hence it follows from  $(r^{k+1}, p^k) = 0$ ,  $(r^{k+1}, r^i) = 0$ ,  $i \leq k-2$ , that  $(r^{k+1}, r^{k-1}) = 0$ . Thus for all  $i \leq k$  one will have  $(r^{k+1}, r^i) = 0$ .  $\square$

If  $r^k$  vanishes, then  $x^k$  is a minimum point of  $f(x)$ . But  $\mathbf{R}^n$  cannot contain more than  $n$  nonzero orthogonal vectors. Hence  $k \leq n$  for some  $r^k = 0$ . Thus we have proven the following theorem.

**THEOREM 2.** Method (13), (16), (17) yields a minimum point of the quadratic function  $f(x)$  (15) in no more than  $n$  iterations.  $\square$

We shall establish in Chapter 7 that if  $L$  is a subspace of  $\mathbf{R}^n$  and  $f(x)$  is a convex differentiable function, then the condition

$$(\nabla f(x^*), a) = 0 \quad \text{for all } a \in L$$

is necessary and sufficient in order that  $x^*$  be a minimum of  $f(x)$  on  $L$ . This and Lemma 1 imply that  $x^k$  is a minimum point of the quadratic function  $f(x)$  (15) on the subspace passing through  $x^0$  and generated by  $r^0, \dots, r^{k-1}$ .

This rather unexpected result (we seek the minimum  $k$  times in succession on 2-dimensional subspaces and find it on the entire  $k$ -dimensional subspace) is an important feature of the conjugate-gradient method thus making its finiteness clear.

The sequential directions  $p^k$  in the conjugate-gradient method satisfy the relation

$$(Ap^i, p^j) = 0, \quad i \neq j. \quad (19)$$

Indeed,  $p^i = x^i - x^{i-1}$ , hence  $Ap^i = Ax^i - Ax^{i-1} = r^i - r^{i-1}$ . On the other hand, we have noted that  $p^k$  is a linear combination of  $r^0, \dots, r^{k-1}$ ,  $p^k = \sum_{j=0}^{k-1} \mu_j r^j$ . Hence for  $i > k$  by Lemma 1 we have

$$(Ap^i, p^k) = \left( r^i - r^{i-1}, \sum_{j=0}^{k-1} \mu_j r^j \right) = 0.$$

Vectors  $p^i$  connected by relation (19) are called *conjugate*, or *A-orthogonal* (they are orthogonal in the metric defined by  $A$ ). This explains the name of the method: conjugate linear combinations of successive gradients are constructed.

Observe that the fact that we know the arbitrary conjugate directions  $s^i$ ,  $i = 1, \dots, n$ ,  $(As^i, s^j) = 0$ ,  $i \neq j$ , allows us to solve easily the system

$$Ax = b, \quad A > 0. \quad (20)$$

Indeed, we will seek the solution in the form  $x = \sum_{i=1}^n \alpha_i s^i$ . Substituting it into (20), computing the scalar product with  $s^i$ , and using the  $A$ -orthogonality, we have

$$\alpha_i = (b, s^i) / (As^i, s^i). \quad (21)$$

This solution can be given a recursive form: we take arbitrary  $x^0$  and construct  $x^k = x^{k-1} + \alpha_k s^{k-1}$ , where the  $\alpha_k$  are given by (21). Then  $x^n = x^*$  is the solution of (20). Since the  $\alpha_k$  in (21) can be determined differently:

$\alpha_k = \underset{\alpha}{\operatorname{argmin}} f(x^{k-1} + \alpha s^k)$ , we see that the fact that we know the system of conjugate directions makes it possible to find the minimum of a quadratic function by means of  $n$  one-dimensional minimizations. This important result will be used repeatedly in what follows in constructing other minimization methods. In the conjugate-gradient method the conjugate directions are not chosen beforehand but constructed from recurrence formulas.

When method (13), (14) is applied to nonquadratic functions, we can easily prove its global convergence if we compare method (13), (14) with the steepest descent method; if we compare it with the heavy-ball method, it is not hard to estimate its convergence rate (Exercises 3 and 4).

The conjugate-gradient method can be given yet another form. Consider the iterative process

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k p^k, & \alpha_k &= \underset{\alpha \geq 0}{\operatorname{argmin}} f(x^k + \alpha p^k), \\ p^k &= -r^k + \beta_k p^{k-1}, & \beta_k &= \|r^k\|^2 / \|r^{k-1}\|^2, \\ r^k &= \nabla f(x^k), & \beta_0 &= 0, \end{aligned} \tag{22}$$

**LEMMA 2.** For the quadratic function (15), methods (13), (16), (17), and (22) for the same  $x^0$  define the same sequence of points  $x^k$ .  $\square$

Since the  $p^k$  in (22) and the  $p^{k+1}$  in (16) differ only by (nonzero) scalar factors, while the  $r^k$  in (22) and (16) coincide, process (22) possesses the same properties as (13), (16): the vectors  $p^i$  are conjugate and the gradients  $r^i$  are mutually orthogonal. Lemma 2 and Theorem 2 imply that method (22) yields a minimum point of the quadratic function (15) in  $\mathbf{R}^n$  in the number of iterations not larger than  $n$ . For nonquadratic problems method (22) is simpler than (13), (14) since it requires solution only of a one-dimensional (rather than a two-dimensional) auxiliary minimization problem. Of course, in the nonquadratic case the finiteness property of the method is lost and (22) turns, in general, into an infinite two-step iterative method. A result concerning its convergence is given in Exercise 5.

For nonquadratic problems the conjugate-gradient method is usually applied in a rather different form, where a restart procedure is introduced: at intervals of time the step is not made by formula (22) but as at the initial point, i.e., according to the gradient. It is most natural to make the restart in terms of the number of iterations equal to the dimension of the space:

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k p^k, & \alpha_k &= \underset{\alpha \geq 0}{\operatorname{argmin}} f(x^k + \alpha p^k), \\ s^k &= -r^k + \beta_k p^{k-1}, & r^k &= \nabla f(x^k), \\ \beta_k &= \begin{cases} 0, & k = 0, n, 2n, \dots \\ \|r^k\|^2 / \|r^{k-1}\|^2, & k \neq 0, n, 2n, \dots \end{cases} \end{aligned} \tag{23}$$

It is not hard to prove that the conjugate gradient method with restart possesses the property of global convergence (Exercise 6). It turns out that it converges, too, with quadratic rate in a neighborhood of the minimum.

**THEOREM 3.** Let  $x^*$  be a nonsingular minimum point and let  $\nabla^2 f(x)$  satisfy a Lipschitz condition in a neighborhood of  $x^*$ . Then for method (23) in a neighborhood of  $x^*$  one has the estimate

$$\|x^{(m+1)n} - x^*\| \leq c \|x^{mn} - x^*\|^2.$$

In other words, with respect to the rate of convergence the  $n$  steps of the conjugate-gradient method are equivalent to one step of Newton's method. We will not give a proof of the theorem since it is rather involved. It is based on the idea of quadratically approximating  $f(x)$  and the fact that the method is finite for quadratic functions (see Theorem 2).  $\square$

Some other computational schemes for the conjugate-gradient method for nonquadratic functions are also possible. We used one of these schemes—requiring solution of a two-dimensional minimization problem on each step—to begin our analysis of this method (see (13), (14)). Other schemes, similarly to (22), usually include only one-dimensional auxiliary problems, but they differ from (22) in the rule for choosing  $\beta_k$ . The scheme

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k p^k, & \alpha_k &= \underset{\alpha \geq 0}{\operatorname{argmin}} f(x^k + \alpha p^k), \\ p^k &= -r^k + \beta_k p^{k-1}, & \beta_k &= \frac{(r^k, r^k - r^{k-1})}{\|r^{k-1}\|^2}, \\ r^k &= \nabla f(x^k), & \beta_0 &= 0, \end{aligned} \tag{24}$$

can serve as an example. Similar to (22), a variant with restart or without restart is possible. For a quadratic function the sequences  $x^k$  generated by methods (22) and (24) coincide.

As numerical computations show, for the nonquadratic case scheme (24) usually gives a slightly faster convergence.

Of interest is the behavior of the conjugate gradient method for large-scale problems (when the number of iterations is smaller than the dimension). It turns out that one can guarantee a convergence with the rate of geometric progression even for the quadratic case. Let  $A$  be an  $n \times n$ -matrix,

$$\ell I \leq A \leq L I, \quad \ell > 0, \tag{25}$$

and let  $f(x)$  be the corresponding quadratic function on  $\mathbf{R}^n$ :

$$f(x) = (Ax, x)/2 - (b, x), \quad b \in \mathbf{R}^n. \tag{26}$$

Then  $x^k$  can be represented in the form

$$x^k - x^* = P_k(A)(x^0 - x^*) ,$$

where  $P_k(A)$  is a matrix polynomial of degree  $k$  of the form

$$P_k(A) = I + a_{1k}A + \cdots + a_{kk}A^k .$$

Thus, the polynomial  $P_k(\lambda)$  satisfies the condition  $2(f(x^k) - f(x^*)) = (AP_k^2(A)(x^0 - x^*), x^0 - x^*) \leq (AR^2(A)(x^0 - x^*), x^0 - x^*)$  where  $R(\lambda)$  is an arbitrary polynomial of degree  $k$  with  $R(0) = 1$  (this follows from the property of  $x^k$  in the conjugate-gradient method to be the minimum point of  $f(x)$  on the subspace passing through  $x^0$  and generated by  $r^0, \dots, r^{k-1}$ ). Hence

$$\begin{aligned} \|x^k - x^*\|^2 &\leq 2(f(x^k) - f(x^*))/\ell \leq \|A\| \|R^2(A)\| \|x^0 - x^*\|^2/\ell \\ &\leq (L/\ell) \|x^0 - x^*\| \sqrt{\max_{\ell \leq \lambda \leq L} R^2(\lambda)} . \end{aligned} \quad (27)$$

Let us choose as  $R(\lambda)$  the polynomial of degree  $k$  with  $R(0) = 1$  having the least absolute deviation from 0 on  $[\ell, L]$ . Such a polynomial is equal to

$$R(\lambda) = T_k \left( \frac{L+\ell-2\lambda}{L-\ell} \right) / T_k \left( \frac{L+\ell}{L-\ell} \right) , \quad (28)$$

where  $T_k(z)$  is the Chebyshev polynomial

$$T_k(z) = \begin{cases} [(z + \sqrt{z^2-1})^k + (z - \sqrt{z^2-1})^k]/2 , & |z| > 1 \\ \cos(k \arccos z) , & |z| \leq 1 . \end{cases} \quad (29)$$

Then

$$\begin{aligned} \max_{\ell \leq \lambda \leq L} R^2(\lambda) &= T_k^{-2} \left( \frac{L+\ell}{L-\ell} \right) \max_{-1 \leq z \leq 1} T_k^2(z) = T_k^{-2} \left( \frac{L+\ell}{L-\ell} \right) \\ &= 4(q^k + q^{-k})^{-2} \leq 4q^{2k} , \quad q = \frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}} . \end{aligned}$$

Hence

$$\|x^k - x^*\| \leq 2 \left( \frac{L}{\ell} \right)^{1/2} q^k \|x^0 - x^*\| , \quad q = (\sqrt{L} - \sqrt{\ell})/(\sqrt{L} + \sqrt{\ell}) . \quad (30)$$

One can show by examples that estimate (30) is unimprovable.

Thus, for  $k < n$  for the conjugate gradient method used to minimize a quadratic function one can guarantee a convergence with the rate of geometric progression with ratio

$$q = (\sqrt{L} - \sqrt{\ell})/(\sqrt{L} + \sqrt{\ell}) \sim 1 - 2\sqrt{\mu}, \quad \mu = L/\ell,$$

i.e., the same as for the heavy-ball method for the optimal choice of its parameters. Versus the latter method, in the conjugate gradient method the choice of parameters presents no problem: they are determined automatically, although they do involve additional computations for solving the one-dimensional minimization problem.

It is obvious that in the conjugate-gradient method the  $x^k$  is a minimum point of the quadratic function  $f(x)$  on the subspace generated by the first  $k$  gradients. It then follows that no method using only gradients of the function (more precisely, the one in which a step is made according to a linear combination of the preceding gradients) can converge more rapidly. In other words, the conjugate gradient method is optimal with respect to its rate of convergence in the class of first-order methods. The result obtained above implies that for large-scale problems with quadratic functions  $f(x)$  satisfying condition (25), for all first-order methods one cannot expect convergence of a higher rate than the rate of geometric progression with ratio  $q = (\sqrt{L} - \sqrt{\ell})/(\sqrt{L} + \sqrt{\ell})$ . Naturally, a higher rate of convergence can neither be attained in the broader class of strongly convex functions with constant  $\ell$ , whose gradient satisfies a Lipschitz condition with constant  $L$ . The quadratic convergence (Theorem 3) occurs only when the number of iterations is significantly greater than the dimension of the space.

## Exercises

2. Check that if  $x'$  is chosen arbitrarily (not by formula (17)), then method (13), (16) converges to the minimum point of (15) with the rate of geometric progression, but it is, in general, not finite. To prove it, one can use, for instance, the fact that by the definition of method (13), (14)  $f(x^{k+1}) \leq f(\bar{x}^{k+1})$ , where  $\bar{x}^{k+1}$  is the point obtained from  $x^k$ ,  $x^{k-1}$  via the heavy-ball method.
3. Let  $f(x)$  be a continuously differentiable function and let the set  $\{x: f(x) \leq f(x^0)\}$  be bounded. Prove that for any  $x^0$ ,  $x^1$  in method (13), (14) one has  $\nabla f(x^k) \rightarrow 0$  (use Theorem 1 of Section 3.1).
4. Let  $x^*$  be a nonsingular minimum point of  $f(x)$ . Following the arguments of Exercise 2, prove the local convergence of method (13), (14) with the rate of geometric progression.

5. Prove the following result about the convergence of the conjugate-gradient method. Let  $f(x)$  be a differentiable strongly convex function whose gradient satisfies a Lipschitz condition. Then method (22) converges for any  $x^0$  to the minimum of  $f(x)$ . Use the following properties of the method:  $(r^k, p^{k-1}) = 0$ ,  $(r^k, p^k) = -\|r^k\|^2$  and the Abel-Dini lemma (the series  $\sum_{k=0}^{\infty} \varepsilon_k$ ,  $\sum_{k=0}^{\infty} \varepsilon_k / (\varepsilon_0 + \dots + \varepsilon_k)$  converge or diverge simultaneously), applying it to  $\varepsilon_k = \|r^k\|^2 / (\beta_1^2 \dots \beta_k^2)$ . Try to estimate the convergence rate.
6. Let  $f(x)$  be continuously differentiable and let the set  $\{x: f(x) \leq f(x^0)\}$  be bounded. Prove that in method (23) one has  $\nabla f(x^k) \rightarrow 0$ . The same holds for any rule of choosing the restart moments if their number is infinite.
7. Prove that  $T_k(\lambda)$  defined by (29) is in fact a polynomial of degree  $k$ .

### 3.3 OTHER FIRST ORDER METHODS

All the methods described in this section are based on the idea of reconstructing a quadratic approximation of a function from values of its gradients at a number of points. Those methods thereby combine the merits of the gradient method (no calculation of the matrix of second derivatives is required) and those of Newton's method (rapid convergence as a result of quadratic approximation).

#### 3.3.1 Quasi-Newton Methods

These methods are generally the following:

$$x^{k+1} = x^k - \gamma_k H_k \nabla f(x^k), \quad (1)$$

where the matrix  $H_k$  is updated recursively on the basis of information obtained in the  $k$ th iteration, so that  $H_k - [\nabla^2 f(x^k)]^{-1} \rightarrow 0$ . Thus in the limit these methods turn into Newton's method, which explains the terminology. Let us note some general properties of such methods. The lemmas below can easily be proved using the technique described earlier.

**LEMMA 1.** Let  $f(x) \geq f^*$ , let  $f(x)$  be differentiable, let  $\nabla f(x)$  satisfy a Lipschitz condition and let

$$mI \leq H_k \leq MI, \quad m > 0. \quad (2)$$

Then in method (1) with  $\gamma_k \equiv \gamma$ , where  $\gamma > 0$  is sufficiently small, one has  $\nabla f(x^k) \rightarrow 0$ .  $\square$

**LEMMA 2.** Let  $x^*$  be a nonsingular minimum point of  $f(x)$ , let  $f(x)$  be twice continuously differentiable in a neighborhood of  $x^*$  and let

$$\|H_k - [\nabla^2 f(x^*)]^{-1}\| \rightarrow 0. \quad (3)$$

Then method (1) with  $\gamma_k = 1$  converges locally to  $x^*$  faster than any geometric progression.  $\square$

Thus, for any uniformly positive definite  $H_k$  method (1) possesses global convergence, and under condition (3) it converges in a neighborhood of the minimum point with superlinear rate.

Let us examine now different ways of constructing matrices  $H_k$  approximating  $[\nabla^2 f(x^k)]^{-1}$ . Theoretically, they can be constructed by finite-difference approximation. Namely, from each point  $x^k$  one can make  $n$  “trial steps” of size  $\alpha_k$  along the coordinates and compute the gradients at these points. The corresponding difference approximation is the one sought if  $\alpha_k \rightarrow 0$  (see Exercise 1).

But such a straightforward method of approximation is inefficient, for it involves  $n$  trial computations of the gradient on each iteration and does not use the gradients obtained in the preceding iterations. The key idea of the quasi-Newton methods is (1) to avoid special trial steps and use, instead, the gradients found at the preceding points (since they are close to  $x^k$ ) and (2) to construct an approximation immediately for the inverse matrix  $[\nabla^2 f(x^k)]^{-1}$ . Let

$$p^k = -H_k \nabla f(x^k), \quad y^k = \nabla f(x^{k+1}) - \nabla f(x^k). \quad (4)$$

Then for the quadratic function  $f(x) = (Ax, x)/2 - (b, x)$ ,  $A > 0$ , we have  $y^k = A(x^{k+1} - x^k) = \gamma_k A p^k$ , i.e.,

$$\gamma_k p^k = A^{-1} y^k. \quad (5)$$

Hence for a new approximation  $H_{k+1}$  of  $[\nabla^2 f(x^{k+1})]^{-1}$  it is natural to require that the so-called *quasi-Newton constraint*

$$H_{k+1} y^k = \gamma_k p^k \quad (6)$$

be satisfied. Furthermore, it is convenient to obtain  $H_{k+1}$  as a correction to  $H_k$  using matrices of rank 1 or 2. Finally, these corrections should be such that for the quadratic case  $H_n = A^{-1}$ .

The basic technique for analyzing such methods is the following lemma on matrix inversion.

**LEMMA 3.** Let  $B$  be an  $n \times n$ -matrix, let  $B^{-1}$  exist, let  $a, b \in \mathbb{R}^n$ . Also, let  $(B^{-1}a, b) \neq -1$ , and let  $A = B + ab^T$ . Then 1(-1)

$$A^{-1} = B^{-1} - (1 + (B^{-1}a, b))^{-1} B^{-1} a \cancel{(B^{-1}b)^T}. \quad (7) \quad H \not\in B^{-1}$$

The lemma is proved by straightforward verification.  $\square$

Thus, if  $B^{-1}$  is known, while  $A$  equals  $B$  plus a rank-one matrix, then  $A^{-1}$  can be found easily.

The following are formulas to update the  $H_k$ .

(a) The Davidon-Fletcher-Powell method (DFP):

$$H_{k+1} = H_k - \frac{H_k y^k (y^k)^T H_k}{(H_k y^k, y^k)} + \gamma_k \frac{p^k (p^k)^T}{(p^k, y^k)}, \quad H_0 > 0; \quad (8)$$

(b) The Broyden method:

$$H_{k+1} = H_k - \frac{(\gamma_k p^k - H_k y^k)(\gamma_k p^k - H_k y^k)^T}{(\gamma_k p^k - H_k y^k, y^k)}, \quad H_0 > 0; \quad (9)$$

(c) The Broyden-Fletcher-Shanno method (BFSH):

$$\begin{aligned} H_{k+1} &= H_k + \frac{\rho_k p^k (p^k)^T - p^k (y^k)^T H_k - H_k y^k (p^k)^T}{(y^k, p^k)}, \\ \rho_k &= \gamma_k + \frac{(H_k y^k, y^k)}{(y^k, p^k)}, \quad H_0 > 0. \end{aligned} \quad (10) \quad \checkmark$$

It turns out that the quasi-Newton constraint (6) holds for each formula (8), (9) or (10). Also, if  $\gamma_k > 0$  are arbitrary,  $p^k$  are arbitrary linearly independent vectors, the  $y^k$  satisfy relation (5) with  $A^{-1} > 0$ , then for any  $H_0 > 0$ ,  $H_n = A^{-1}$ . This implies the following theorem.

**THEOREM 1.** For any  $x^0, H_0 > 0$  method (1), (4) with any of the updating formulas (8), (9) or (10) and  $\gamma_k = \underset{\gamma}{\operatorname{argmin}} f(x^k + \gamma p^k)$  for  $f(x) = (Ax, x)/2 - (b, x)$ ,  $A > 0$ , is finite:  $x^n = x^* = A^{-1}b$ .  $\square$

Furthermore, one can show that regardless the differences between the updating formulas the sequences  $x^k$  generated by each variant of the method coincide for a quadratic function  $f(x)$ .

For nonquadratic functions the quasi-Newton methods in the form given above are usable, but they are no longer finite. Therefore, for  $k > n$  one can either continue the computation by the same formulas or begin a restart procedure (replacing  $H_k$  by  $H_0$  every  $n$  iterations).

Currently a superlinear (or quadratic) rate of convergence has been proved for many variants of quasi-Newton methods in a neighborhood of a nonsingular minimum point.

These results seem natural in terms of Lemmas 1 and 2 and Theorem 1, but their complete proof is very cumbersome.

Quasi-Newton methods are widely used and have been extensively treated in the literature, due to the numerous advantages as we described earlier: the computation of the gradient at each step; no matrix inversion, nor solution of a system of linear equations; global convergence; a high rate of convergence in a neighborhood of the solution (often quadratic rate), among others. Yet they are inferior, say, to the conjugate-gradient method: the need to store and update an  $n \times n$ -matrix  $H_k$  with significant computer storage for large  $n$  is the greatest disadvantage.

Variant (10) of the quasi-Newton methods usually yields the best results in numerical verification of the methods.

### Exercise

- Let  $e_1, \dots, e_n$  be the coordinate basis vectors in  $\mathbf{R}^n$ , let  $f(x)$  be differentiable in a neighborhood of  $x$  and twice differentiable at  $x$ . Let  $H(\alpha)$  be the matrix with  $\alpha^{-1}(\nabla f(x + \alpha e_i) - \nabla f(x))$  as the  $i$ th row. Show that  $H(\alpha) \rightarrow \nabla^2 f(x)$  as  $\alpha \rightarrow 0$ .

### 3.3.2 Methods of Variable Metric and Methods of Conjugate Directions

We derived the quasi-Newton methods as approximations to Newton's method. They can, however, be interpreted differently.

First of all, let us see how the choice of the metric affects the form and the properties of the gradient method. Suppose that in the space  $\mathbf{R}^n$  in addition to the initial scalar product  $(x, y)$  a scalar product defined by a matrix  $A > 0$  is given:

$$(x, y)_1 = (Ax, y). \quad (11)$$

In this case the  $A$  defines a new metric in  $\mathbf{R}^n$ :

$$\|x - y\|_1^2 = (A(x - y), x - y). \quad (12)$$

Let us write the gradient of a differentiable function  $f(x)$  in the new metric:

$$\begin{aligned} f(x+y) &= f(x) + (\nabla f(x), y) + o(\|y\|) = f(x) + (AA^{-1}\nabla f(x), y) + o(\|y\|) \\ &= f(x) + (a, y)_1 + o(\|y\|_1), \\ a &= A^{-1}\nabla f(x). \end{aligned}$$

By definition, the vector  $a$  is the gradient of  $f(x)$  in space with scalar product (11). Thus,

$$\nabla_1 f(x) = A^{-1}\nabla f(x). \quad (13)$$

In the new metric the gradient method assumes the form

$$x^{k+1} = x^k - \gamma_k \nabla_1 f(x^k) = x^k - \gamma_k A^{-1}\nabla f(x^k) \quad (14)$$

and differs from the original gradient method by the presence of the matrix  $A^{-1}$ . In other words, the gradient method is not invariant with respect to the choice of metric of the space. It is reasonable to choose the metric such as to increase the rate of convergence. For the quadratic function

$$f(x) = (Bx, x)/2 - (b, x) = (\frac{1}{2})(A^{-1}Bx, x)_1 - (A^{-1}b, x)_1 \quad (15)$$

the convergence rate of (14) is determined by the progression ratio  $q = (L - \ell)(L + \ell)$ , where  $L$  and  $\ell$  are respectively the largest and the smallest eigenvalues of  $A^{-1}B$ . The closer the  $A^{-1}B$  to the unit matrix, the smaller  $q$ . The best way is to choose  $A = B$ , because then  $A^{-1}B = I$ ,  $q = 0$ , i.e., if one defines the metric with the matrix  $B$ , then the gradient method (with  $\gamma_k \equiv 1$ ) will yield an accurate solution in one step. This is not surprising, for  $f(x) = (1/2)(x, x)_1 - (A^{-1}b, x)_1$ , i.e., the level lines of the  $f(x)$  are spheres and the condition number  $\mu$  is equal to one.

For a nonquadratic function the method

$$x^{k+1} = x^k - \gamma_k H_k \nabla f(x^k), \quad H_k > 0, \quad (16)$$

can be viewed as the gradient method in the metric

$$(x, y)_1 = (H_k^{-1}x, y), \quad (17)$$

and  $H_k = [\nabla^2 f(x^k)]^{-1}$  is the “optimal” choice of the metric. In other words, the quasi-Newton methods can be treated as gradient methods in which

a new metric is chosen on each step as close to the best one as possible. For this reason the term *methods of a variable metric* is often synonymous to that of quasi-Newton methods.

This interpretation is also useful as a heuristic construction of new variants of quasi-Newton methods. For example, one can obtain a new metric by extending the space in the direction of the last gradient, or in the direction of the difference of two consecutive gradients, and the like. We will discuss these methods in more detail in Chapter 5.

Yet another approach to constructing efficient first-order methods involves the notion of conjugate directions. As was observed in Section 3.2, the knowledge of the set of conjugate directions  $p^0, \dots, p^{n-1}$ :

$$(Ap^i, p^j) = 0, \quad i \neq j, \quad (18)$$

makes it possible to find the minimum of a quadratic function  $f(x) = (Ax, x)/2 - (b, x)$  in  $n$  one-dimensional minimizations:

$$x^{k+1} = x^k + \alpha_k p^k, \quad \alpha_k = \underset{\alpha}{\operatorname{argmin}} f(x^k + \alpha p^k). \quad (19)$$

Then  $x^* = x^* = A^{-1}b$  for any  $x^0$ . One of the methods for constructing conjugate directions was used in the conjugate gradient method: the sequentially computed gradients were subjected to the  $A$ -orthogonalization. Other methods are quite possible as well.

Let  $p^1, \dots, p^k, k < n-1$  be conjugate vectors that have been constructed,

$$(Ap^i, p^j) = 0, \quad 0 \leq i, j \leq k, \quad i \neq j, \quad (20)$$

and let  $x^k$  be the corresponding points in method (19). The next vector  $p^{k+1}$  must satisfy the relation

$$(p^{k+1}, Ap^i) = 0, \quad i = 0, \dots, k.$$

Since

$$p^i = \alpha_i^{-1}(x^{i+1} - x^i), \quad Ap^i = \alpha_i^{-1}(\nabla f(x^{i+1}) - \nabla f(x^i)) = \alpha_i^{-1}y^i,$$

this is equivalent to the condition

$$(p^{k+1}, y^i) = 0, \quad i = 1, \dots, k. \quad (21)$$

Thus, the new conjugate direction  $p^{k+1}$  must satisfy the orthogonality conditions (21). Orthogonalization of any linearly independent vectors gives us varied sets of conjugate directions.

The same process can be used for a nonquadratic function:

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k p^k, \quad \alpha_k = \underset{\alpha \geq 0}{\operatorname{argmin}} f(x^k + \alpha p^k) \\ (p^{k+1}, y^i) &= 0, \quad i = 1, \dots, k, \quad y^i = \nabla f(x^{i+1}) - \nabla f(x^i). \end{aligned} \quad (22)$$

Usually,  $p^{k+1}$  is sought here in the form

$$p^{k+1} = -H_{k+1} \nabla f(x^{k+1}), \quad H_{k+1} = H_k + \Delta H_k \quad (23)$$

and the matrix  $H_k$  is stored instead of the vectors  $y^i$ ,  $i = 1, \dots, k$ . The methods thus assume the same form (1) as the quasi-Newton methods. The only difference is that it is not necessarily  $H_k \rightarrow [\nabla^2 f(x^k)]^{-1}$ ; in some variants of the method  $H_n = 0$  (for a quadratic function). That is why in these methods one must use a restart procedure.

We will next write an algorithm for one of the simplest methods of this class:

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k p^k, \quad \alpha_k = \underset{\alpha \geq 0}{\operatorname{argmin}} f(x^k + \alpha p^k), \\ p^k &= -H_k \nabla f(x^k), \quad y^k = \nabla f(x^{k+1}) - \nabla f(x^k), \\ H_{k+1} &= H_k - \frac{H_k y^k (y^k)^T H_k}{(H_k y^k, y^k)}, \quad k+1 \neq n, 2n, \dots, \\ H_0 &= H_n = H_{2n} = \cdots = I. \end{aligned} \quad (24)$$

It turns out that for a quadratic function in method (24) the  $p^k$  are conjugate directions,  $H_k \geq 0$  for all  $k \leq n$ ,  $H_n = 0$ . For nonquadratic functions the quadratic local convergence of methods of this class in a neighborhood of a nonsingular minimum point has been proved.

### 3.3.3 The Secant Method

One the simplest and most commonly used methods for solving the one-dimensional equation

$$g(x) = 0 \quad (25)$$

is the secant method illustrated in Figure 7. This method can be extended to the multidimensional case: if  $g: \mathbf{R}^n \rightarrow \mathbf{R}^n$ , then one can compute  $g$  at  $n+1$  points, construct a linear approximation and find its root which is the closest approximation to the solution of (25).

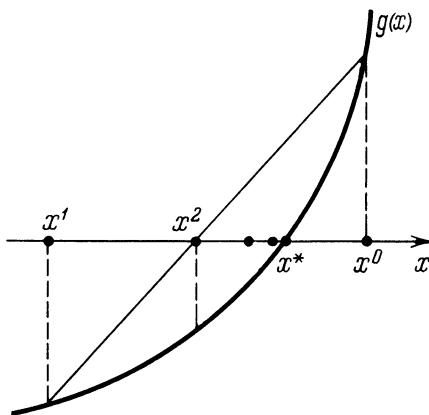


Fig. 7 The secant method.

For the problem of minimizing  $f(x)$  on  $\mathbf{R}^n$ , i.e., the problem of solving the equation  $\nabla f(x) = 0$ , the secant method becomes the following. Let  $x^k, x^{k-1}, \dots, x^{k-n}$  be  $n+1$  points in  $\mathbf{R}^n$ , and let  $\nabla f(x^k), \dots, \nabla f(x^{k-n})$  be the gradients computed at these points. We solve the system of  $n+1$  linear equations with  $n+1$  variables  $\lambda_0, \lambda_1, \dots, \lambda_n$ :

$$\sum_{i=0}^n \lambda_i \nabla f(x^{k-i}) = 0, \quad \sum_{i=0}^n \lambda_i = 1. \quad (26)$$

Also, let us construct the point

$$x^{k+1} = \sum_{i=0}^n \lambda_i x^{k-i}. \quad (27)$$

Then the process is repeated for the last  $n+1$  points  $x^{k+1}, x^k, \dots, x^{k-n+1}$  and so on. It is not hard to check that for  $n = 1$  this method coincides with the secant method for solving the equation  $\nabla f(x) = 0$ .

**THEOREM 2.** If the vectors  $x^1 - x^0, x^2 - x^0, \dots, x^n - x^0$  are linearly independent and  $f(x)$  is quadratic with  $\nabla^2 f(x) \equiv A > 0$ , then  $x^{n+1}$  is the minimum point of  $f(x)$ .  $\square$

In the system of linear equations (26), only one column changes in each iteration, and therefore there is no need to solve it each time—one might use as well the following lemma.

**LEMMA 4.** Let  $B$  be a (square)  $n \times n$ -matrix with columns  $b^1, \dots, b^n$ . Also, let  $\tilde{B}$  differ from it only in the first column (say,  $b^1$  is replaced by  $\tilde{b}^1$ ). Then

$$\tilde{c}^i = c^i - \frac{(\tilde{b}^1 - b^1, c^i)}{1 + (\tilde{b}^1 - b^1, c^i)} c^1, \quad (28)$$

where  $c^i$  is the row of  $B^{-1}$ ,  $\tilde{c}^i$  is the row of  $\tilde{B}^{-1}$ .

To prove the lemma it suffices to represent  $\tilde{B}$  in the form  $\tilde{B} = B + (\tilde{b}^1 - b^1)e^T$ , where  $e = (1, 0, \dots, 0)$ , and use Lemma 3.  $\square$

However, the secant method written in this form is not adequate, viz. it does not have the property of global convergence. To eliminate this drawback, one can proceed in the standard way, for example, adjust the step size (from  $x^k$  the step is made in the direction  $\sum_i \lambda_i x^{k-i}$ ). Another drawback is that the method has a tendency to degenerate: during the computations the sequential approximations lie (approximately) in a subspace of  $\mathbf{R}^n$ . The corresponding system of linear equations (26) is ill-conditioned and its solution is unstable. To overcome this drawback, one can modify the method so as to make the system of basis points *a priori* nonsingular. For example, one can add one point at a time in each iteration by making a step along the coordinates (in cyclic order). For such augmented methods one can prove superlinear convergence.

### 3.3.4 Other Approaches for Constructing the First-order Methods

Regardless the variety of first-order algorithms the idea behind them was the same for all of them, viz. to use a quadratic approximation of the function near the minimum. As a rule, these algorithms are finite for quadratic functions and in the general case they are more efficient if their function is closer to being quadratic. But the quadratic model can be regarded to be natural only in a neighborhood of the extremum; far from the extremum the behavior of the objective function may be somewhat different. Hence for all of the methods it is clearly not advisable to apply an optimization strategy even at the initial stages of the search.

Instead, it is advantageous to use models of functions other than quadratic. It seems natural to make an attempt to construct polynomial models using higher derivatives: the next terms of the Taylor series. This has been tried before—however without good results. First, a direct computation of higher derivatives in multidimensional problems is usually too cumbersome and requires large memory; furthermore, to reconstruct them from lower derivatives one needs to compute them at a too large number of points. Secondly, auxiliary problems of minimizing polynomial functions cannot, with rare exception, be solved in the analytic form.

A simple and important class of models includes those based on the approximation of a homogeneous function. The function  $f(x)$ ,  $x \in \mathbf{R}^n$ , is called *homogeneous* with respect to  $x^*$  with exponential  $\gamma > 0$  if

$$f(x^* + \lambda(x - x^*)) - f(x^*) = \lambda^\gamma(f(x) - f(x^*)) \quad (29)$$

for all  $x \in \mathbf{R}^n$  and  $\lambda \geq 0$ . Examples of homogeneous functions are given in Exercises 2, 3, 4 and 6.

A differentiable homogeneous function satisfies the important relation

$$f(x) - f(x^*) = \gamma^{-1}(\nabla f(x), x - x^*) . \quad (30)$$

To prove (30) we take  $\lambda = 1 + \varepsilon$  in (29):

$$\begin{aligned} \text{V}^- ) \quad f(x + \varepsilon(x - x^*)) - f(x^*) &= (1 + \varepsilon)^\gamma(f(x) - f(x^*)) , \\ \varepsilon\gamma(f(x) - f(x^*)) &= \varepsilon(\nabla f(x), x - x^*) + o(\varepsilon) . \end{aligned}$$

Letting  $\varepsilon$  go to zero yields (30).

The point  $x^*$  is not necessarily a minimum point of  $f(x)$  (see the examples in Exercises 2 and 3). However, if  $f(x)$  attains a minimum, then  $x^*$  is a global minimum point of  $f(x)$ . Indeed, let  $f(\bar{x}) = f^* = \min f(x)$ . Then  $\nabla f(\bar{x}) = 0$ . Substituting  $\bar{x}$  for  $x$  into (30), we get  $f(x^*) = f(\bar{x}) = f^*$ , i.e.,  $x^*$  is a global minimum point. We shall be examining this particular case later.

Using (30), one can find the minimum point  $x^*$  through computation of  $f(x)$  and  $\nabla f(x)$  at a finite number of points. Indeed, if  $\gamma$  is known, then taking  $n+1$  points  $x^0, \dots, x^n$ , yields the system

$$\gamma f(x^i) - \alpha + (\nabla f(x^i), x^*) = (\nabla f(x^i), x^i) , \quad i = 0, \dots, n , \quad (31)$$

which is linear in the  $n+1$  variables  $x^*, \alpha$  ( $\alpha$  ( $\alpha = \gamma f(x^*)$ )). Eliminating  $\alpha$ , we obtain  $n$  linear equations to determine  $x^* \in \mathbf{R}^n$ :

$$\begin{aligned} (\nabla f(x^i) - \nabla f(x^0), x^*) &= (\nabla f(x^i), x^i) - (\nabla f(x^0), x^0) - \gamma(f(x^i) - f(x^0)) , \\ i &= 1, \dots, n . \end{aligned} \quad (32)$$

But if  $\gamma$  is unknown, then one can take  $n+2$  points  $x^0, \dots, x^{n+1}$ , and determine the  $n+1$  variables  $\gamma, x^*$  from the linear system (32) in which  $n+1$  equations have to be taken.

A similar approach can be applied to minimize functions of general form, as was done in the secant method. Indeed, let the approximations  $x^0, \dots, x^k$ ,  $k > n$  have been constructed. Taking the last  $n+1$  among them (or  $n+2$ ) if  $\gamma$  is unknown), we solve the system (with respect to  $x, \alpha, \gamma$ , or  $x, \alpha$ )

$$(\nabla f(x^i) - \alpha + \gamma f(x^i)) = (\nabla f(x^i), x^i), \quad i = k, k-1, \dots, \quad (33)$$

and for  $x^{k+1}$  take the solution  $x$ . For  $\gamma = 2$  we get a method similar to the secant method but not exactly the same—unlike the secant method, the method obtained uses both  $\nabla f(x^i)$  and the values of the function  $f(x^i)$ .

Such a process should be modified using the same techniques as for the secant method (for example, eliminating the degeneration of points  $x^k$  by adding new points which are linearly independent of the preceding points; or adjusting the step size). A comparison of the actual value of  $f(x^{k+1})$  with the “predicted” value (equal to  $\alpha/\gamma$ ) is also useful to verify the assumption concerning the proximity of the function to being homogeneous. In solving systems of linear equations it is appropriate to take advantage of the closeness of these equations in successive iterations (see Lemma 4).

To minimize homogeneous functions or functions close to being homogeneous, some other methods can be used. For example, in the gradient method one uses special techniques for choosing the step size. Let the function  $f(x)$  satisfy condition (30), with the  $f^* = f(x^*)$  and  $\gamma$  being known. We consider the gradient method

$$x^{k+1} = x^k - \frac{\gamma(f(x^k) - f^*)}{\|\nabla f(x^k)\|^2} \nabla f(x^k). \quad (34)$$

The step

$$\gamma_k = \frac{\gamma(f(x^k) - f^*)}{\|\nabla f(x^k)\|^2}$$

is chosen such that the equality  $f(x^k) - f^* = \gamma^{-1}(\nabla f(x^k), x^k - x^{k+1})$  is satisfied for  $x^{k+1} = x^k - \gamma_k \nabla f(x^k)$  (cf. (30)). Then

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - x^*\|^2 - \frac{2\gamma(f(x^k) - f^*)}{\|\nabla f(x^k)\|^2} (\nabla f(x^k), x^k - x^*) \\ &\quad + \frac{\gamma^2 (f(x^k) - f^*)^2}{\|\nabla f(x^k)\|^2} \\ &= \|x^k - x^*\|^2 - \frac{\gamma^2 (f(x^k) - f^*)^2}{\|\nabla f(x^k)\|^2} \end{aligned}$$

implying that if  $\|\nabla f(x)\|$  is bounded on the set  $\{x: \|x - x^*\| \leq \|x^0 - x^*\|\}$ , then  $f(x^k) \rightarrow f^*$ . It is not hard to see that this result still holds if in (30) equality is replaced by inequality

$$f(x) \neq f^* \leq \gamma^{-1}(\nabla f(x), x - x^*). \quad (35)$$



A somewhat different class (versus the homogeneous one) is given by the formula

$$f(x) = F(\phi(x)) , \quad \phi(x) = (Ax, x)/2 - (b, x) , \quad A > 0 , \quad (36)$$

where  $F: \mathbf{R}^1 \rightarrow \mathbf{R}^1$  is a monotone function on  $[\phi^*, \infty]$ ,  $\phi^* = \phi(x^*)$ . Obviously,  $x^*$  is a minimum point of  $f(x)$ .

If  $F$  and  $\phi$  are given in the explicit form, a simpler problem of minimizing  $\phi(x)$  can be solved instead of the problem of minimizing  $f(x)$ . In general, however, the information on the problem is not sufficient. Then the following variant of the conjugate-gradient method can be used:

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k p^k , & \alpha_k &= \underset{\alpha \geq 0}{\operatorname{argmin}} f(x^k + \alpha p^k) , \\ p^k &= -\nabla f(x^k) + \beta_k p^{k-1} , & & \\ \beta_k &= \frac{F'(\phi(x^{k-1})) \| \nabla f(x^k) \|^2}{F'(\phi(x^k)) \| \nabla f(x^{k-1}) \|^2} , & \beta_0 &= 0 . \end{aligned} \quad (37)$$

It is not hard to check that method (37) generates the same sequence of points as the conjugate-gradient method does for minimization of  $\phi(x)$ ; it is therefore finite.

The quantity  $\rho_k = F'(\phi(x^k))/F'(\phi(x^{k-1}))$  in the formula for  $\beta_k$  can be estimated approximately via approximation of the  $F(z)$  by a quadratic or a power function. In that case method (37) can be used to minimize functions that do not necessarily have the form (36).

On the whole, methods based on homogeneous approximations of functions have not been studied with adequate thoroughness.

### Exercises

2. Show that the affine function  $f(x) = (a, x) - \beta$  is homogeneous with  $\gamma = 1$  for any  $x^*$ .
3. Verify that the quadratic function  $f(x) = (Ax, x)/2 - (b, x)$ , where  $A^{-1}$  exists, is homogeneous with respect to  $x^* = A^{-1}b$  with  $\gamma = 2$ .
4. Suppose there exists a solution  $x^*$  of the system  $(a^i, x) = \beta_i$ ,  $i = 1, \dots, m$ ,  $x \in \mathbf{R}^n$ . Prove that the function

$$f(x) = \sum_{i=1}^m |(a^i, x) - \beta_i|^\gamma , \quad \gamma > 0 ,$$

is homogeneous with respect to  $x^*$  with exponent  $\gamma$ .

5. Prove that for a twice-differentiable homogeneous function the relation  $\nabla^2 f(x)(x-x^*) = (\gamma-1)\nabla f(x)$  holds.
6. Show that if  $\phi^* = 0$  and  $F(z) = |z|^\alpha$ ,  $\alpha > 0$ , then  $f(x)$  of the form (36) is homogeneous with respect to  $x^*$  with exponent  $2\alpha$ .

## 3.4 DIRECT METHODS

### 3.4.1 General Characteristics

In many problems the objective function is given by an algorithm for computing its values at an arbitrary point. The form of the algorithm may be unknown (for example, the values of the function are determined either by means of the model or on the real system). Or, the algorithm can be complex so that the analytic computation of the gradient is too involved. In all such cases the values of the function  $f(x)$  are the only available information. Methods which employ the information on  $f(x)$  only are called *zero-order methods* (often referred to as *direct methods*, *search methods* or *methods without derivatives*).

The most straightforward strategy in these situations consists in using the values of the function for a finite-difference approximation of the derivatives—the gradient or the Hessian. A most efficient method takes account of the values of the function at the preceding points. Also, there are several special zero-order methods; they have no analogs among first- or second-order methods.

### 3.4.2 Methods of Linear Approximation

To estimate the gradient of  $f: \mathbf{R}^n \rightarrow \mathbf{R}$  at a point  $x$ , we form finite-difference relations

$$\Delta_1 = \alpha^{-1} [f(x+\alpha y) - f(x)] , \quad \Delta_2 = (2\alpha)^{-1} [f(x+\alpha y) - f(x-\alpha y)] , \quad (1)$$

where  $y \in \mathbf{R}^n$  is an arbitrary vector.

**LEMMA 1.** (a) If  $f$  is differentiable at  $x$ , then

$$|\Delta_1 - (\nabla f(x), y)| \rightarrow 0 \quad \text{as } \alpha \rightarrow 0 . \quad (2)$$

(b) If  $\nabla f$  satisfies a Lipschitz condition with constant  $L$  in a neighborhood of  $x$ , then for a sufficiently small  $\alpha$

$$|\Delta_1 - (\nabla f(x), y)| \leq L \alpha \|y\|^2 / 2 . \quad (3)$$

(c) If  $f$  is twice differentiable and  $\nabla^2 f$  satisfies a Lipschitz condition in a neighborhood of  $x$ , then for a sufficiently small  $\alpha$

$$|\Delta_2 - (\nabla f(x), y)| \leq L \alpha^2 \|y\|^3/6. \quad (4)$$

(d) If  $f(x)$  is quadratic, then for any  $\alpha$

$$\Delta_2 = (\nabla f(x), y). \quad (5)$$

Lemma 1 is easily proved if one uses (2), (15), (20) of Section 1.1.  $\square$

Thus, the difference relations  $\Delta_1$  and  $\Delta_2$  may serve as an approximation for linear approximation of  $f(x)$ . Let us consider methods of the form

$$x^{k+1} = x^k - \gamma_k s^k, \quad (6)$$

where  $\gamma_k \geq 0$  is the step size and  $s^k$  is computed by one of these two formulas:

$$s^k = \sum_{i=1}^m \alpha_k^{-1} [f(x^k + \alpha_k h^i) - f(x^k)] h^i, \quad (7)$$

$$s^k = \sum_{i=1}^m (2\alpha_k)^{-1} [f(x^k + \alpha_k h^i) - f(x^k - \alpha_k h^i)] h^i, \quad (8)$$

where  $h^i$ ,  $i = 1, \dots, m$ , are vectors giving the directions of the trial steps,  $\alpha_k$  is the size of the trial step. By choosing  $h^i$  and  $m$  we obtain various algorithms.

(a) The difference analog of the gradient method:  $m = n$ ,  $h^i = e_i$ ,  $i = 1, \dots, n$ , where the  $e_i$  are the standard basis vectors. In other words, the trial steps are made along the coordinates so that method (6), (7) has the following form in coordinate notation:

$$x_i^{k+1} = x_i^k - (\gamma_k / \alpha_k) [f(x^k + \alpha_k e_i) - f(x^k)]. \quad (9)$$

By Lemma 1

$$s^k = \sum_{i=1}^n (\nabla f(x^k), e_i) e_i + \varepsilon^k = \nabla f(x^k) + \varepsilon^k, \quad (10)$$

where the remainder  $\varepsilon^k$  can be estimated either for (7) or for (8), depending on the smoothness of  $f(x)$ .

(b) Method of coordinatewise descent:  $m = 1$ ,  $h = e_j$ ,  $j = k(\text{mod } n)$ .

The steps are made along the coordinates chosen in cyclic order:

$$x_i^{k+1} = \begin{cases} x_i^k - (\gamma_k/\alpha_k) [f(x^k + \alpha_k e_i) - f(x^k)], & i = k \pmod{n}, \\ x_i^k & \text{otherwise.} \end{cases} \quad (11)$$

Here  $s^k = \nabla f(x^k); e_j + \epsilon^k$ .

(c) Method of random coordinatewise descent:  $m = 1$ ,  $h = e_j$ , where  $j$  takes on the values  $1, \dots, n$  equiprobably. The step is made as above along the coordinates, but they are chosen in random order.

(d) Method of random search:  $m = 1$ ,  $h$  is a random vector uniformly distributed on the unit sphere. The direction of the step is random, the sign and the size of the step are determined by the difference relation

$$x^{k+1} = x^k - (\gamma_k/\alpha_k) [f(x^k + \alpha_k h) - f(x^k)] h. \quad (12)$$

The convergence of all the methods is guaranteed by the condition  $\alpha_k \rightarrow 0$  (see Exercise 1).

The rate of convergence depends on the smoothness of  $f(x)$  and the choice of  $\alpha_k$ . To minimize errors, it is more convenient to take large  $\alpha_k$  since the smaller the  $\alpha_k$  the greater the effect of roundoff errors in computing difference relations (in (1) one needs to compute the difference between two close quantities and then divide by a small number; of course this causes loss in accuracy). However, for large  $\alpha_k$  the accuracy of approximation is worse (Lemma 1). One can show that by hypothesis of Theorem 3 of Section 1.4, in the methods described above the convergence with the rate of geometric progression is guaranteed if  $\alpha_k \leq cq^k$ , where  $q \neq 1$  is an integer.

The question of the convergence rates in the respective methods is a difficult one. We consider first a special case which may serve as a model for more realistic situations. Let  $f(x)$  be quadratic:

$$f(x) = (Ax, x)/2 - (b, x), \quad A > 0, \quad (13)$$

and let  $\gamma_k$  be chosen from the steepest descent condition:

$$x^{k+1} = x^k - \gamma_k s^k, \quad \gamma_k = \underset{\gamma \geq 0}{\operatorname{argmin}} f(x^k - \gamma s^k). \quad (14)$$

We shall compare three methods of choosing the  $s^k$ : (1) the symmetric difference approximation of the gradient

$$s^k = \sum_{i=1}^n (2\alpha)^{-1} [f(x^k + \alpha e_i) - f(x^k - \alpha e_i)] e_i = \nabla f(x^k) \quad (15)$$

(the last equality is due to (5)); (2) coordinatewise descent

$$s^k = (2\alpha)^{-1} [f(x^k + \alpha e_i) - f(x^k - \alpha e_i)] e_i = \nabla f(x^k)_i e_i, \quad i = k(\text{mod } n) \quad (16)$$

and (3) random search

$$\underbrace{h^k}_{s^k} = (2\alpha)^{-1} [f(x^k + \alpha h^k) - f(x^k - \alpha h^k)] h^k = (\nabla f(x^k), h^k) h^k, \quad (17)$$

where  $h^k$  is a vector uniformly distributed on the unit sphere. Thus, method (14), (15) coincides with the steepest descent method ((4) of Section 3.1), while method (14), (16) is well known in Linear Algebra as the Gauss-Seidel method.

The correspondence between the methods and the rate of convergence is a function of many factors. Here are a few special, extreme cases. If  $A = I$ , then method (14), (15) and method (14), (16) lead to solution in one step, whereas the random-search method converges in mean square no quicker than some geometric progression. If  $(Ax, x) = \sum_{i=1}^n \lambda_i x_i^2$ ,  $\lambda_i > 0$ , then method (14), (16) is finite, whereas method (14), (15) is not. If the problem is ill-posed ( $\mu \gg 1$ ), one can show that the random-search method converges more rapidly than the gradient method (taking into account the difference in the number of computations of  $f(x)$  in one iteration of each method). Roughly, for such problems the random direction is a better indication of the solution than the antigradient direction. The Gauss-Seidel method has another additional lane of increasing the convergence: if  $\gamma_k$  is replaced by  $\alpha\gamma_k$ ,  $1 < \alpha < 2$  (the so-called overrelaxation), the convergence improves in a number of cases.

To conclude, among the search methods of this class the method of coordinatewise descent is superior to other methods for its simplicity and the rate of convergence.

### Exercises

1. Prove that by the hypothesis of Theorem 1 of Section 1.4, as  $\alpha_k \rightarrow 0$ ,  $\gamma_k = \gamma$ , where  $\gamma$  is sufficiently small, one can assert for all the methods a-d that  $\nabla f(x^k) \rightarrow 0$  a.s. Use the technique employed in proving Theorem 1 of Section 2.2.
2. Suggest a constructive method for regulating  $\alpha_k$  to guarantee the linear rate of convergence, by analogy with (10) of Section 3.1.

### 3.4.3 Nonlocal Linear Approximation

In the finite-difference gradient method (9) the trial and the operational steps were distinct, i.e., the points  $x^k + \alpha_k e_i$  served only for estimation of the gradient at  $x^k$ , whereas at  $x^{k+1}$  the procedure repeats. One

can proceed differently and construct a linear approximation from the set of points at sufficient intervals.

The so-called *simplicial method* (not to be confused with the simplex method in linear programming!) is a typical example. Suppose there are  $n+1$  points  $x^0, x^1, \dots, x^n$ , being the vertices of a regular simplex. We compute the values of  $f(x)$  at the vertices and find the vertex at which  $f(x)$  is maximal:  $j = \underset{0 \leq i \leq n}{\operatorname{argmax}} f(x^i)$ . Next, we construct a new simplex different from the old one only in one vertex:  $x^j$  is replaced by  $x^{n+1}$ :

$$x^{n+1} = 2n^{-1}(x^0 + \dots + x^{j-1} + x^{j+1} + \dots + x^n) - x^j \quad (18)$$

(i.e.,  $x^{n+1}$  is symmetric to  $x^j$  with respect to the opposite side). If it turns out that the maximum is attained at the vertex  $x^{n+1}$  in the new simplex, we go back to the initial simplex, replacing  $x^j$  by the vertex at which the value of  $f(x)$  is maximal versus the remaining vertices, etc. If some point remains in  $n+1$  successive simplices, the last simplex is reduced to one half by a similarity transformation centered at this vertex (Fig. 8).

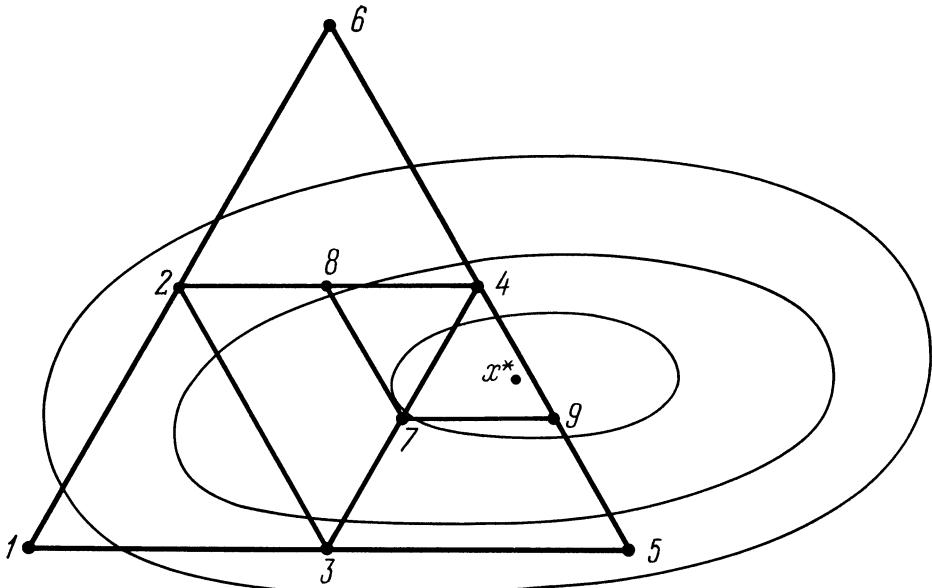


Fig. 8 The simplicial method.

In addition to this simplest variant of the simplicial method, there exist many modifications in which the simplex need not be regular, and the step size and the conditions for subdivision are different. These methods have not been investigated thoroughly enough in theoretical terms. As practice showed, they are good enough when the problems are not too ill-posed.

### 3.4.4 Quadratic Approximation

Using the values of  $f(x)$  at sufficiently many points, one can construct the quadratic approximation of  $f(x)$ . For example, the *method of barycentric coordinates* can be used for this purpose. As in the simplicial method, one chooses  $n+1$  basis points  $x^0, \dots, x^n$ . Then one computes the values of the  $f(x)$  at all of these points and at the midpoints on the segments joining them (let  $f((x^i+x^j)/2) = f_{ij}$ ,  $f(x^i) = f_{ii}$ ,  $i, j = 0, \dots, n$ ). Lastly, one solves the system of linear (with respect to  $\lambda, \lambda_0, \dots, \lambda_n$ ) equations

$$\begin{aligned} 4 \sum_{j=0}^n f_{ij} \lambda_j + \lambda &= f_{ii}, \quad i = 0, \dots, n, \\ \sum_{j=0}^n \lambda_j &= 1 \end{aligned} \tag{19}$$

and constructs the point

$$x^{n+1} = \sum_{i=0}^n \lambda_i x^i. \tag{20}$$

It is not hard to verify that if  $f$  is quadratic, then  $x^{n+1} = x^* = A^{-1}b$  for *any*  $x^0, \dots, x^n$  such that  $x^n - x^0, \dots, x^1 - x^0$  are linearly independent.

Next (for a nonquadratic  $f(x)$ ) one includes the point  $x^{n+1}$  into the basis points and removes one of the old basis points ( $x^0$  or the point at which  $f(x)$  is maximal). In the next successive iteration it is sufficient to compute  $f(x)$  at  $n+1$  points (at  $x^{n+1}$  and the midpoints of the segments joining  $x^{n+1}$  with the other basis points). The new system of equations for  $\lambda_i$  will differ from (19) by one row only, so that one can employ the result of Lemma 4 in Section 3.3 to construct the solution. The process proceeds in the similar manner.

The advantage of this method is the fact that one does not write explicitly the actual quadratic approximation of the function but constructs only the minimum point of this approximation. Compared with the finite-difference analog of Newton's method, the method of barycentric coordinates requires essentially smaller amount of computations of  $f(x)$  at each step ( $n+1$  instead of  $n(n+1)/2$ ). To give stability to the process, one has to make adjustments for the step size, prevent the degeneration of the

system of basis points, verify the convexity condition  $f_{ij} \leq (f_{ii} + f_{jj})/2$ , and so on.

Another group of methods of direct search are based on the ideas of the method of conjugate directions and reduce the initial problem to a sequence of one-dimensional minimizations. In contrast with the method of coordinatewise descent, in which the system of descent directions (coordinate orts) is rigidly fixed, in the methods of this group the descent directions are constructed in the process of minimization. To construct them means to make them (for the problem of minimizing a quadratic function) conjugate; then (see Section 3.2) the minimization process is finite in the quadratic case. The key idea of these methods is illustrated in Figure 9: three successive one-dimensional minimizations lead to the minimum point. A similar result holds true in the multidimensional case as well.

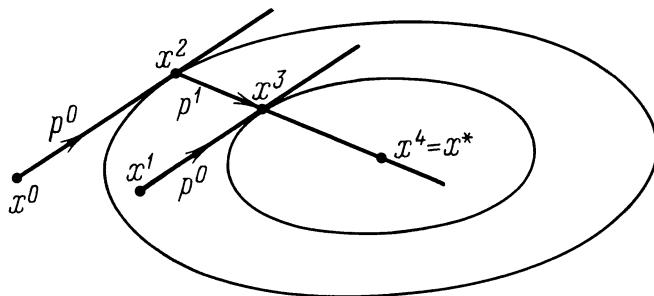


Fig. 9 The method of conjugate directions.

**LEMMA 2.** Let  $f(x) = (Ax, x)/2 - (b, x)$ ,  $A > 0$ ,  $x \in \mathbb{R}^n$ ,  $p^1, \dots, p^k$  be conjugate vectors:

$$(Ap^i, p^j) = 0, \quad i \neq j, \quad k < n,$$

$$L^0 = \left\{ x: x = x^0 + \sum_{i=1}^k \lambda_i p^i \right\}, \quad x^1 \in L^0,$$

$$L^1 = \left\{ x: x = x^1 + \sum_{i=1}^k \lambda_i p^i \right\},$$

$$y^0 = \underset{x \in L^0}{\operatorname{argmin}} f(x), \quad y^1 = \underset{x \in L^1}{\operatorname{argmin}} f(x).$$

Then the vector  $p^{k+1} = y^1 - y^0$  is conjugate to  $p^1, \dots, p^k$ .

This lemma follows from the condition for a minimum of  $f(x)$  on a subspace (see the remark following Theorem 2 in Section 2.2).  $\square$

One can thus construct a minimization method in the following way. Let  $x^k$  be the approximation to the solution obtained in the  $k$ th iteration, and let  $p^0, \dots, p^k$  be the resulting directions ( $x^0$  and  $p^0$  are arbitrary). We construct  $\bar{x}^k = x^k + h^k$ ,  $h^k$  is an arbitrary vector not being a linear combination of  $p^0, \dots, p^k$ . Now let us make sequential one-dimensional minimizations in the directions  $p^0, \dots, p^k$ , starting at the point  $\bar{x}^k$ ; we obtain the point  $\hat{x}^{k+1}$ . For the point  $x^{k+1}$  we take the minimum point of  $f(x)$  on the line joining  $\hat{x}^{k+1}$  with  $x^k$ , and for  $p^{k+1}$  we take the vector  $\hat{x}^{k+1} - x^k$ . For a quadratic function in  $\mathbf{R}^n$ , this method—the Powell method—leads to the minimum in no more than  $n$  steps.

There are many other modifications based on the same idea. To find the minimum in the quadratic case one needs  $n(n+1)/2$  minimizations. If each minimization includes, say, three computations of the function, the method of barycentric coordinates is clearly less efficient than method (19), (20) (which requires  $n(n+1)/2$  minimizations for the same purpose). However in the nonquadratic case the Powell method is efficient enough even in case of a poor initial approximation (one prevents the degeneration of the system of  $p^i$ ), whereas the method of barycentric coordinates, as well as Newton's method, requires a good initial approximation.

## CHAPTER 4

### INFLUENCE OF NOISE

In this chapter our objective is to observe the behavior of methods of unconstrained minimization for differential functions in the presence of noise. It has been proved that the methods have different sensitivity to noise, i.e., the more effective the method is in the ideal situation (without noise), the more sensitive it is to different kinds of errors. One can modify the methods so as to make them operable in the presence of noise. In this case the *a priori* information about the noise (the level, the distribution law, etc.) can be very useful.

#### 4.1 SOURCES AND TYPES OF NOISE

##### 4.1.1 Sources of Noise

In real problem the methods described in Chapters 1 and 3 cannot be applied in “pure form” because of the unavoidable errors and inaccuracies. We shall explain some of the reasons for them.

In the simplest case where the objective function and its gradient are given by formulas, inaccuracies are the result of computational errors due to roundoff in arithmetical operations on a computer. As a result,  $f(x^k)$ ,  $\nabla f(x^k)$ , and the like, are computed with some error, i.e., instead of the vector  $\nabla f(x^k)$  we obtain the vector  $s^k = \nabla f(x^k) + r^k$ . Here the noise  $r^k$  is deterministic (the computer roundoff errors are not of a random nature) and its level  $\|r^k\| \leq \varepsilon$ , can be estimated since the laws concerning the occurrence of roundoff errors have been studied thoroughly enough. The variable  $\varepsilon$  can be usually assumed to be constant (not depending on  $x^k$ ) and

generally not too large. If necessary,  $\varepsilon$  can be made smaller by making calculations with double precision.

In some problems the values of  $f(x^k)$  and  $\nabla f(x^k)$  obtain not through computations but by means of measurements. This is observed in the optimization of a real system (extremal control, experiment design). In that case noise is random, which is characteristic of measurement errors; however, the information about the level as well as the statistical nature of the noise is usually available to the user.

In problems of adaptation, learning, pattern recognition, among others, the optimization problem is usually the following. It is required to minimize the deterministic function  $f(x)$  of mean risk type:

$$f(x) = \mathbf{E}Q(x, \omega) = \int Q(x, \omega) d\mathbf{P}(\omega), \quad (1)$$

where the function  $Q(x, \omega)$  is known but the distribution  $\mathbf{P}(\omega)$  is not specified. Only a sample  $\omega_1, \dots, \omega_k$  of  $\mathbf{P}(\omega)$  is given. Then the exact computation of  $f(x)$  and  $\nabla f(x)$  is, in principle, impossible. As an approximate value of these variables one can take

$$\frac{1}{k} \sum_{i=1}^k Q(x, \omega_i) \quad \text{and} \quad \frac{1}{k} \sum_{i=1}^k \nabla_x Q(x, \omega_i), \quad (2)$$

or more simply

$$Q(x, \omega_k) \quad \text{and} \quad \nabla_x Q(x, \omega_k). \quad (3)$$

In this case the values of the function and gradient contain a random noise. If one takes  $Q(x^k, \omega_k)$  and  $\nabla_x Q(x^k, \omega_k)$  as approximations for  $f(x^k)$  and  $\nabla f(x^k)$ , then the noise at different points is mutually independent.

A similar situation arises in the *Monte-Carlo method*, in which the problem consists in minimization of  $f(x)$  of the form (1), the distribution  $\mathbf{P}(\omega)$  is known, but the computation of the integral (1) is too involved. In this case the exact values of  $f(x)$  and  $\nabla f(x)$  can be replaced by sample values, as above.

In some problems errors obtain because the values of the function or the gradient are computed by too simple or too approximate formulas. Frequently, exact computation requires an elaborate computation of influence functions, solution of complex auxiliary problem, the interaction of all of the parameters, and the like. It is not recommended (sometimes even impossible) to make complete computations. A simplification or coarsening of these computations leads to inaccuracies in determining the function and the gradient. These are known as *unavoidable errors*.

Finally, in many methods errors occur due to the need of solving auxiliary problems, which cannot be done with precision. For example, in Newton's method, at each step one needs to solve a system of linear equations,

and this always involves errors; in the conjugate-gradient method it is required to make a one-dimensional minimization, which can be done only approximately, etc. These are known as *errors of the method*.

#### 4.1.2 Types of Noise

As was shown earlier, errors in computing the function and gradient can have different origin and nature. In general, the basic types of noise are the following. Everywhere in the sequel we shall be dealing with a computation of the gradient when instead of the exact value of  $\nabla f(x^k)$  we have the vector

$$s^k = \nabla f(x^k) + r^k, \quad (4)$$

where  $r^k$  is the noise. A case of approximate computation of  $f(x)$  can be investigated in a similar way (see Section 4.4).

(a) *Absolute deterministic noise* satisfies the condition

$$\|r^k\| \leq \varepsilon, \quad (5)$$

i.e., the gradient is computed with a given absolute error. It is assumed that nothing except this condition is known about the noise. In particular, the vector  $r^k$  need not be random, or it can be correlated with the preceding noise, and so on. Such a situation is typical for computational errors and systematic measurement errors.

(b) *Relative deterministic noise* satisfies the condition

$$\|r^k\| \leq \varepsilon \|\nabla f(x^k)\|. \quad (6)$$

In other words, the gradient is calculated with a relative error. As above, nothing except this condition is known about the nature of the  $r^k$ . Such noise occurs, for example, in using approximate formulas involving a fixed relative error.

(c) *Absolute random noise*. Suppose that the noise  $r^k$  is random, independent for different  $x$ , centered and has bounded variance:

$$\mathbf{E}r^k = 0, \quad \mathbf{E}\|r^k\|^2 \leq \sigma^2. \quad (7)$$

Such noise is typical for problems in which the gradient is being sought through measurements of a real system (extremal control, experimental design), and also for problems with mean risk function (1).

(d) *Relative random noise* possesses the same properties as in (c). However, the noise variance decreases as it approaches a minimum point:

$$\mathbf{E}r^k = 0, \quad \mathbf{E}\|r^k\|^2 \leq \tau \|\nabla f(x^k)\|^2. \quad (8)$$

Of course, other types of noise, too, are observed in practice; for example, random noise with systematic error ( $\|Er^k\| \leq \varepsilon$ ) or random bounded noise ( $Er^k = 0$ ,  $\|r^k\| \leq \varepsilon$ ). But they can be treated as a combination of the types listed above. Hence we shall limit ourselves to an examination of these, most important classes of noise. Sometimes (especially in theoretical works) it is assumed that the level of noise  $\varepsilon_k$  depends on the number of the iteration and that  $\varepsilon_k \rightarrow 0$  as  $k \rightarrow \infty$ . Such an assumption does not seem to be realistic. Nevertheless in some cases it may hold if the computations have been made more accurate, thus decreasing the error of the method.

## 4.2 THE GRADIENT METHOD IN THE PRESENCE OF NOISE

### 4.2.1 The Statement of the Problem

Let us consider the gradient method for minimizing the differentiable function  $f(x)$  on  $\mathbf{R}^n$  in the situation when the gradient computed with error:

$$x^{k+1} = x^k - \gamma_k s^k, \quad s^k = \nabla f(x^k) + r^k. \quad (1)$$

In respect to the noise  $r^k$ , we assume that it belongs to one of the classes described in Section 4.1. The function  $f(x)$  is assumed to be strongly convex (with constant  $\ell$ ) and with a gradient satisfying a Lipschitz condition (with constant  $L$ )—this class of functions is most important (see Chapters 1 and 3). We are interested in the behavior of the ordinary gradient method with  $\gamma_k \equiv \gamma$  in the presence of noise, as well as the choice of the step size. We shall prove these methods, using the general theorems of Section 2.2.

### 4.2.2 Absolute Deterministic Noise

**THEOREM 1.** Let  $\|r^k\| \leq \varepsilon$ ,  $\gamma_k \equiv \gamma$ . Then we can find a  $\bar{\gamma} > 0$  such that for  $0 < \gamma < \bar{\gamma}$  in method (1) we have

$$\|x^k - x^*\| \leq \rho + q^k \|x^0 - x^*\|, \quad (2)$$

where  $0 \leq q < 1$ ,  $\rho = \rho(\varepsilon) \rightarrow 0$  as  $\varepsilon \rightarrow 0$ ,  $x^*$  is the minimum point of  $f(x^k)$ .

**PROOF.** Introduce the Lyapunov function

$$V(x) = \frac{1}{2} \left( \|x - x^*\| - \frac{1}{\ell} \varepsilon \right)_+^2. \quad (3)$$

Using the result of Exercise 1 below, we obtain

$$\begin{aligned} (\nabla V(x^k), s^k) &= \left( \|x^k - x^*\| - \frac{1}{\ell} \varepsilon \right)_+ \frac{(\nabla f(x^k) + r^k, x^k - x^*)}{\|x^k - x^*\|} \\ &\geq \left( \|x^k - x^*\| - \frac{1}{\ell} \varepsilon \right)_+ (\ell \|x^k - x^*\| - \varepsilon) = 2\ell V(x^k), \end{aligned}$$
↗

$$\begin{aligned} \|s^k\|^2 &= \|\nabla f(x^k) + r^k\|^2 \leq (L \|x^k - x^*\| + \varepsilon)^2 \leq a + bV(x^k) \\ &\leq a + (b/(2\ell))(\nabla V(x^k), s^k), \end{aligned}$$

where  $a$  and  $b$  are constants, and  $a \rightarrow 0$  as  $\varepsilon \rightarrow 0$ . Applying Theorem 4 of Section 2.2 proves the theorem.  $\square$

It is not hard to verify by using examples (see Exercise 2 below) that estimate (2) is not overrated. Thus, the presence of additive noise leads to the situation that the gradient method no longer converges with constant  $\gamma$  to a minimum point. It creates the possibility, however, that the method might get in a neighborhood of the minimum, the size of which depends on the noise level. The method converges to this neighborhood with the rate of geometric progression.

We did not write the exact values of the constants  $\rho$ ,  $\gamma$ ,  $q$  since we are interested only in the qualitative evaluation of the process. In Exercise 2 below these values are given for a case of a quadratic function.

## Exercises

1. Prove that  $V(x)$  of the form of (3) is differentiable,  $\nabla V(x) = (\|x - x^*\| - \varepsilon/\ell)_+ \times \|x - x^*\|^{-1}(x - x^*)$ ,  $\nabla V(x)$  satisfies a Lipschitz condition with constant 1. Sketch the graph of  $V(x)$  for  $x \in \mathbf{R}^1$ .

2. Let

$$f(x) = (Ax, x)/2 - (b, x), \quad \ell I \leq A \leq L I, \quad \ell > 0, \quad \|r^k\| \leq \varepsilon, \quad 0 < \gamma < 2/L.$$

Show that in method (1) one has

$$\|x^{k+1} - x^*\| \leq q \|x^k - x^*\| + \gamma \varepsilon, \quad q = \max \{|1 - \gamma \ell|, |1 - \gamma L|\}.$$

Using Lemma 2 of Section 2.2, derive the estimate

$$\|x^k - x^*\| \leq \gamma \varepsilon / (1 - q) + q^k (\|x^0 - x^*\| - \gamma \varepsilon / (1 - q)).$$
↙ 1

In particular, for  $\gamma = 2/(L + \ell)$ , it then follows that

$$\|x^k - x^*\| \leq \frac{\varepsilon}{\ell} + \left( \|x^0 - x^*\| - \frac{\varepsilon}{\ell} \right) \left( \frac{L-\ell}{L+\ell} \right)^k.$$

Verify by using an example that this estimate is not overrated. Investigate the limit case  $\varepsilon = 0$ .

### 4.2.3 Relative Deterministic Noise

**THEOREM 2.** Let

$$\|r^k\| \leq \alpha \|\nabla f(x^k)\|, \quad \alpha < 1, \quad \gamma_k \equiv \gamma.$$

Then we can find a  $\bar{\gamma} > 0$  such that for  $0 < \gamma < \bar{\gamma}$  method (1) converges to  $x^*$  with the rate of geometric progression.

**PROOF.** Take  $V(x) = f(x) - f(x^*)$  as a Lyapunov function. Then (see Lemmas 1 and 3 in Section 1.4) we have

$$\begin{aligned} (\nabla V(x), s^k) &= (\nabla f(x^k), \nabla f(x^k) + r^k) \geq (1 - \alpha) \|\nabla f(x^k)\|^2 \\ &\geq (1 - \alpha) 2\ell V(x^k), \\ \|s^k\|^2 &\leq \|\nabla f(x^k)\|^2 (1 + \alpha)^2 \leq 2(1 + \alpha)^2 L V(x^k). \end{aligned}$$

It remains only to apply Theorem 4 of Section 2.2.  $\square$

Thus the gradient method is stable under relative errors if their level is less than 100%. This is obvious: any direction that makes an acute angle with the antigradient is the direction of decrease of  $f(x)$  and may be used as a direction of motion instead of the gradient.

### 4.2.4 Absolute Random Noise

Let the noise  $r^k$  be random, independent, and let  $E r^k = 0$  and  $E \|r^k\| \leq \sigma^2$ .

**THEOREM 3.** We can find a  $\bar{\gamma} > 0$  such that for  $\gamma_k \equiv \gamma$ ,  $0 < \gamma < \bar{\gamma}$ , in method (1) we have

$$E(f(x^k) - f^*) \leq \rho(\gamma) + E(f(x^0) - f^*) q^k, \quad (4)$$

where  $q < 1$ ,  $\rho(\gamma) \rightarrow 0$  as  $\gamma \rightarrow 0$ . If

$$\gamma_k \rightarrow 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty, \quad (5)$$

then  $E \|x^k - x^*\|^2 \rightarrow 0$ . If though

$$\sum_{k=0}^{\infty} \gamma_k^2 < \infty, \quad \sum_{k=0}^{\infty} \gamma_k = \infty, \quad (6)$$

then  $x^k \rightarrow x^*$  a.s. Finally, if  $\gamma_k = \gamma/k$ ,  $\gamma > (2\ell)^{-1}$ , then

$$E(f(x^k) - f^*) \leq \frac{L\sigma^2\gamma^2}{2(2\ell\gamma - 1)k} + o\left(\frac{1}{k}\right). \quad (7)$$

**PROOF.** Take  $V(x) = f(x) - f^*$ . Then

$$\begin{aligned} (\nabla V(x^k), Es^k) &= (\nabla f(x^k), \nabla f(x^k)) \geq 2\ell V(x^k), \\ E \|s^k\|^2 &= \|\nabla f(x^k)\|^2 + E \|r^k\|^2 \leq \sigma^2 + (\nabla V(x^k), Es^k). \end{aligned}$$

It remains only to use Theorems 2-5 of Section 2.2.  $\square$

We shall see later (Theorem 4) that the foregoing estimates are not overrated, and hence Theorem 3 permits the following conclusions. First, the usual variant of the gradient method (with  $\gamma_k \equiv \gamma$ ) in the presence of additive random noise does not converge to a minimum point but, rather, leads to a neighborhood of the minimum. The smaller  $\gamma$ , the smaller the size of this region. Secondly, choosing decreasing  $\gamma_k$  may make the method converge in some probabilistic sense (in the mean as  $\gamma_k \rightarrow 0$  or almost surely for  $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$ ). Thirdly, the convergence rate is quite slow in this case (of order  $O(1/k)$ ). As will be shown later, no choice of  $\gamma_k$  can yield a better convergence rate.

Let us refine Theorem 3 for a quadratic function and constant noise level. Thus, let

$$\begin{aligned} f(x) &= (Ax, x)/2 - \underbrace{(f, x)}, \quad \ell I \leq A \leq LI, \quad \ell > 0, \\ Er^k &= 0, \quad Er^k(r^k)^T = \sigma^2 I. \end{aligned} \quad (8) \quad \text{T } \ell$$

We assume that the initial approximation  $x^0$  is random and symmetrically distributed around  $x^*$ :  $E(x^0 - x^*)(x^0 - x^*)^T = \alpha I$ .

**THEOREM 4.** For any  $0 < \gamma < 2/L$ ,  $\gamma_k \equiv \gamma$ , in method (1) under conditions (8) for the quantity

$$U_k = \mathbf{E}(x^k - x^*)(x^k - x^*)^T \quad (9)$$

we have the relations

$$U_k \rightarrow U_\infty = \gamma\sigma^2 A^{-1} (2I - \gamma A)^{-1}, \quad (10)$$

$$\|U_k - U_\infty\| \leq \|U_0 - U_\infty\| q^k, \quad (11)$$

$$q = \max \{(1-\gamma\ell)^2, (1-\gamma L)^2\} < 1.$$

$$\begin{aligned} \text{If } \gamma_k &\not\equiv \gamma/k, \gamma > (2\ell)^{-1} \\ U_k &= \frac{1}{k} B(\gamma) + o\left(\frac{1}{k}\right), \quad B(\gamma) = \gamma\sigma^2 \left[2A - \frac{1}{\gamma} I\right]^{-1}. \end{aligned} \quad (12)$$

The quantity  $\|B(\gamma)\|$  is minimal for  $\gamma = 1/\ell$ ,

$$\|U_k\| = \frac{1}{k} \frac{\sigma^2}{\ell^2} + o\left(\frac{1}{k}\right). \quad \square \quad (13)$$

#### 4.2.5 Relative Random Noise

Let the noise  $r^k$  be as in the previous subsection, but assume that the variance satisfies the condition

$$\mathbf{E}\|r^k\|^2 \leq \alpha\|\nabla f(x)\|^2. \quad (14)$$

**THEOREM 5.** For any  $\alpha$  we can find a  $\bar{\gamma}$  such that for  $0 < \gamma < \bar{\gamma}$ , in method (1) we have

$$\mathbf{E}\|x^k - x^*\|^2 \leq cq^k, \quad q < 1. \quad \square \quad (15)$$

We see that the presence of random relative noise of any level does not lead to violation of the convergence.

Thus, the type of noise determines whether the noise retains or violates the convergence of the gradient method. In some cases the convergence can be renewed by adjusting the step size.

## 4.3 OTHER MINIMIZATION METHODS IN THE PRESENCE OF NOISE

### 4.3.1 Newton's Method

The behavior of Newton's method in the presence of noise is substantially more complicated compared with the gradient method. The reason is that this method may include several sources of noise (computation of  $\nabla f(x)$ ,  $\nabla^2 f(x)$ , inversion of  $\nabla^2 f(x)$ ), and their nature varies—for instance, random errors in computing the gradient and systematic errors in matrix inversion. We make no attempt to cover all possible situations. Only a few typical examples will be considered since we are interested only in the qualitative view of the process.

As a result of all calculations (of the gradient, the Hessian, and solving the system of linear equations), suppose we have a vector differing from the true one:

$$s^k = [\nabla^2 f(x^k)]^{-1} \nabla f(x^k) + r^k, \quad (1)$$

where  $r^k$  is the noise, and the step is

$$x^{k+1} = x^k - s^k. \quad (2)$$

Suppose that the noise contains a systematic error:

$$\|r^k\| \leq \varepsilon. \quad (3)$$

As is known, Newton's method converges locally in some region  $U$ . If  $\varepsilon$  is larger than the diameter of  $U$ , there is no convergence, since for any  $x^0$  arbitrarily close to  $x^*$  the process exits from  $U$ . Thus, as will not happen with the gradient method, Newton's method may behave erratically (for example,  $\|x^k - x^*\|$  may increase) for any  $x^0$  if the noise level is sufficiently high.

Systematic errors in Newton's method are unavoidable even if  $\nabla f(x)$  and  $\nabla^2 f(x)$  are computed precisely, for if the condition number  $\mu$  of the minimum point (Section 1.3) is large (and it is precisely then that it is most expedient to apply Newton's method), the matrix  $\nabla^2 f(x^k)$  is most likely to be ill-conditioned. This leads to the situation that a solution of the system of linear equations  $\nabla^2 f(x^k)z = \nabla f(x^k)$  for determining the step of the method is not an exact solution, due to roundoff errors in the computer. This difference (for ill-conditioned systems) can be significant and may cause Newton's method to fail.

Random or relative errors need not be so catastrophic, but can produce a substantial slowdown in Newton's method. For example, let us minimize

the quadratic function

$$f(x) = (Ax, x)/2 - (b, x), \quad A > 0, \quad (4)$$

where  $A$  and  $A^{-1}$  are computed exactly, and where the gradient contains a random error:

$$s^k = \nabla f(x^k) + r^k = Ax^k - b + r^k, \quad Er^k = 0, \quad E\|r^k\|^2 = \sigma^2. \quad (5)$$

Consider the method

$$x^{k+1} = x^k - \gamma_k A^{-1} s^k, \quad (6)$$

a generalization of Newton's method, by introduction of the parameter  $\gamma_k$ . As we will see later (Theorem 1, Section 4.5), this method cannot converge faster than  $O(1/k)$  for any  $\gamma_k$ . This cancels the basic advantage of Newton's method, that is, its quick convergence rate; the much simpler gradient method can guarantee a convergence of the same order as this generalization. Relative error has a similar situation: if the gradient is calculated with relative error, then Newton's method can only converge with geometric progression rate.

Only highly accurate calculations allow Newton's method to retain its superiority (see Exercise 1).

### Exercise

1. Prove the following result. Let  $r^k$  in (1) satisfy the condition

$$\|r^k\| \leq c \|\nabla f(x^k)\|^2, \quad (7)$$

and let Theorem 1 in Section 1.5 be applicable to  $f(x)$ . Then for sufficiently small  $c$ , method (2) converges locally with quadratic rate.

### 4.3.2 Multistep Methods

We shall again limit our attention to analyzing typical special cases. To begin, it can be seen that under absolute deterministic noise in determining the gradient, the heavy-ball method converges to a neighborhood around the minimum. Cumbersome computation shows that the size of this region is generally greater for a quadratic function than for the gradient method. We can give an analogous result pertaining to random noise. Let

$$\begin{aligned} f(x) &= (Ax, x)/2 - (b, x), & \ell I \leq A \leq L I, \quad \ell > 0, \\ s^k &= \nabla f(x^k) + r^k = Ax^k - b + r^k, & Er^k = 0, \quad E(r^k)^T = \sigma^2 I, \end{aligned} \quad (8)$$

where the  $r^k$  are mutually independent. As can be shown, the heavy-ball method with constant coefficients

$$x^{k+1} = x^k - \alpha s^k + \beta(x^k - x^{k-1}) \quad (9)$$

does not converge to  $x^* = A^{-1}b$  under these conditions, but only leads into a region around  $x^*$ . Hence we will consider the method with variable coefficients, which may be conveniently written as

$$x^{k+1} = x^k - \alpha_k y^k, \quad y^{k+1} = y^k - \beta_k(y^k - s^k). \quad (10)$$

At the same time, let us consider the gradient method

$$x^{k+1} = x^k - \gamma_k s^k, \quad (11)$$

limiting coefficients to the form

$$\alpha_k = \frac{1}{k}\alpha, \quad \beta_k = \frac{1}{k}\beta, \quad \gamma_k = \frac{1}{k}\gamma. \quad (12)$$

**THEOREM 1.** For any  $\alpha, \beta$ , method (10), (12) converges asymptotically no faster (in the sense of the quantity  $\|\mathbf{E}(x^k - x^*)(x^k - x^*)^T\|$ ) than method (11) with  $\gamma_k = 1/k\beta$ !  $\square$

Thus the heavy-ball method is relatively less effective under noise than the gradient method, although it has a faster convergence rate in noise-free problems.

This conclusion pertains only to asymptotic behavior of the method. In early iterations when the relative value of noise is small, the two-step method may exceed the one-step method, as it does in noise-free problems.

The situation is roughly the same for the conjugate gradient method. A full analysis of its behavior under noise is very complicated since different variants of it react differently to errors. Apparently formulas (13), (14) of Section 3.2 are most stable; formulas (23) and (24) of Section 3.2 are somewhat less so. One can show that under absolute and relative noise the conjugate gradient method loses its superiority over the gradient method near the minimum. Only if the noise satisfies a condition like (7) does the conjugate gradient method retain its advantages.

### 4.3.3 Other Methods

Quasi-Newton methods are very sensitive to errors in calculating the gradient. Indeed, in them the matrix  $A = \nabla^2 f(x)$  is restored from measure-

ments of the gradient:

$$\begin{aligned} Ap^i &\approx y^i, \quad p^i = x^{i+1} - x^i, \quad y^i = \nabla f(x^{i+1}) - \nabla f(x^i), \\ i &= 0, \dots, k-1. \end{aligned} \quad (13)$$

If the steps are small ( $x^{i+1}$  is close to  $x^i$ ), and the measurements of  $\nabla f(x^i)$  contain errors, then the matrix is restored poorly. For problems with random additive noise, this can be changed by increasing the number of measurements. It is necessary to make the restoration not from  $n$  values of  $\nabla f(x)$ , as in the deterministic case, but from  $N > n$  measurements. Here one can write out recurrent formulas analogous to those in Section 3.3. For nonrandom noises, this procedure does not generally enhance accuracy.

The secant method has analogous conditions: to make it effective under random noise, the number of base points must be taken as notably larger than the dimension of the space.

However, it must be remembered that the possibilities of all methods based on quadratic approximation are very limited in problems with noise. Even knowing the precise matrix of second derivatives does not save the day (see the analysis for Newton's method in Section 4.2).

## 4.4 DIRECT METHODS

### 4.4.1 The Statement of the Problem

At an arbitrary point  $x^k$  let the value of  $f(x^k)$  be measured with error  $\eta_k$ . As above, we will speak of an absolute (relative) deterministic error if  $|\eta_k| \leq \varepsilon$  ( $|\eta_k| \leq \alpha(f(x^k) - f(x^*))$ ), and of an absolute (relative) random error if the  $\eta_k$  are random, independent,

$$E\eta_k = 0 \quad \text{and} \quad E\eta_k^2 \leq \sigma^2 \quad (E\eta_k^2 \leq \tau(f(x^k) - f(x))).$$

The problem is to study the influence of such errors on primary methods of minimization (Section 3.4) and to modify these methods to overcome the effect of noise.

### 4.4.2 Difference Methods for Random Noise

Let us consider methods of the type given in Section 3.4, in examples with random noise. We will begin with the most typical of these, the *Keifer-Wolfowitz method* (the method of difference approximation of the gradient):

$$x^{k+1} = x^k - \gamma_k s^k, \quad s^k = \sum_{i=1}^n \frac{1}{2\alpha_k} (\tilde{f}(x^k + \alpha_k e_i) - \tilde{f}(x^k - \alpha_k e_i)) e_i, \quad (1)$$

where  $e_i$  are the standard basis vectors. Here and later

$$\tilde{f}(x) = f(x) + \eta , \quad (2)$$

and the random errors  $\eta$  are independent at different points and

$$E\eta = 0 , \quad E\eta^2 \leq \sigma^2 . \quad (3)$$

Let us discuss the trial and working steps  $\alpha_k, \gamma_k$ . Set

$$s^k - \nabla f(x^k) = g^k + \xi^k ,$$

where  $g^k$  is the systematic error, and  $\xi^k$  is the random error. If  $f(x)$  is twice differentiable, and  $\nabla^2 f(x)$  satisfies a Lipschitz condition, then by Lemma 1 of Section 3.4,

$$\|g^k\| \leq c\alpha_k^2 . \quad (4)$$

The random component of the error in estimating the gradient is:

$$E\xi^k = 0 , \quad E\|\xi^k\|^2 \leq \sigma^2/(2\alpha_k^2) . \quad (5)$$

Thus the systematic error decreases as  $\alpha_k$  decreases, but the random error increases. First let us show that  $\alpha_k, \gamma_k$  can be regulated to guarantee convergence.

**THEOREM 1.** Let  $f(x)$  be strongly convex and twice differentiable, let  $\nabla^2 f(x)$  satisfy a Lipschitz condition, let condition (3) hold and for  $\gamma_k, \alpha_k$  let the following relations hold:

$$\begin{aligned} \sum_{k=0}^{\infty} \gamma_k &= \infty , \quad \sum_{k=0}^{\infty} \gamma_k \alpha_k^4 < \infty , \quad \sum_{k=0}^{\infty} \gamma_k^2 \alpha_k^2 < \infty , \\ &\sum_{k=0}^{\infty} \gamma_k^2 \alpha_k^{-2} < \infty . \end{aligned} \quad (6)$$

Then in method (1),  $x^k \rightarrow x^*$  a.s. and  $E\|x^k - x^*\|^2 \rightarrow 0$ . If  $\gamma_k = \gamma/k$ ,  $\alpha_k = \alpha k^{-1/6}$  and  $\gamma$  is sufficiently large, then

$$E\|x^k - x^*\|^2 = O(k^{-2/3}) . \quad \square$$

An analogous result can be derived for a nonsymmetric difference approximation of the gradient under less stringent smoothness assumptions on  $f(x)$  (see Exercise 2).

Thus for convergence under additive random noise in measuring the function it is necessary that both the trial and the working steps tend to 0, and the trial steps must decrease more slowly. The asymptotic convergence rate depends on the choice of  $\alpha_k$ ,  $\gamma_k$ , the smoothness of  $f(x)$  and the form of the difference approximation; however it does not exceed  $O(k^{-s})$ ,  $s < 1$ . These very same conclusions also hold for the more general algorithms in Section 3.4.

Let us give more precise estimates of the convergence rate for a quadratic function under constant additive noise:

$$\begin{aligned} f(x) &= (Ax, x)/2 - (b, x), \quad A \geq \ell I > 0, \quad x \in \mathbf{R}^n, \\ \tilde{f}(x) &= f(x) + \eta, \quad \mathbf{E}\eta = 0, \quad \mathbf{E}\eta^2 = \sigma^2, \end{aligned} \tag{7}$$

where the noise  $\eta$  is independent at different points. Let us compare the (gradient) *Kiefer-Wolfowitz method*

$$\begin{aligned} x^{k+1} &= x^k - \gamma_k s^k, \\ s^k &= \sum_{i=1}^n \frac{1}{2\alpha_k} [\tilde{f}(x^k + \alpha_k e_i) - \tilde{f}(x^k - \alpha_k e_i)] e_i \end{aligned} \tag{8}$$

and the *random search method*

$$\begin{aligned} x^{k+1} &= x^k - \gamma_k s^k, \\ s^k &= (2\alpha_k)^{-1} [\tilde{f}(x^k + \alpha_k h^k) - \tilde{f}(x^k - \alpha_k h^k)] h^k, \end{aligned} \tag{9}$$

where  $h^k$  is a random vector uniformly distributed on the unit sphere (and not depending on  $\eta$ ). Since for a quadratic function the systematic error in the difference approximation of the gradient is equal to 0 for any  $\alpha_k$  (Lemma 1 of Section 3.4), it is not necessary here to make  $\alpha_k$  tend to 0. We shall assume that in (8) and (9)  $\alpha_k \equiv \alpha > 0$ . Using Theorem 4 of Section 4.2, it is not hard to prove that in method (8) for  $\gamma_k = \gamma/k$ ,  $\gamma > 1/(2\ell)$ ,

$$\mathbf{E}(x^k - x^*)(x^k - x^*)^T = \frac{1}{k} \frac{\gamma\sigma^2}{2\alpha^2} \left[ 2A - \frac{1}{\gamma} I \right]^{-1} + o\left(\frac{1}{k}\right), \tag{10}$$

while in method (9) for  $\gamma_k = \gamma/k$ ,  $\gamma > n/(2\ell)$ ,

$$\mathbf{E}(x^k - x^*)(x^k - x^*)^T = \frac{1}{k} \frac{\gamma\sigma^2}{2\alpha^2} \left[ 2A - \frac{n}{\gamma} I \right]^{-1} + o\left(\frac{1}{k}\right). \tag{11}$$

It follows that if  $\gamma_k$  in (8) is taken  $n$  times more than in (9), then  $n$  steps of (9) will be asymptotically equivalent to one step of (8). Noting that the laboriousness of (8) is  $n$  times greater than that of (9), we find that in the present situation (8) and (9) are equivalent in terms of asymptotic efficiency. It is worth mentioning that this conclusion does not depend on the condition number or any other properties of  $A$  (compare with a different situation in noise-free problems in Section 3.4).

Note in conclusion that asymptotic estimates of the kind given in Theorem 1 have to be handled circumspectly. For example, the choice  $\alpha_k = \alpha k^{-1/6}$  means that it is necessary to make a million iterations in order to diminish the trial step by a factor of ten. Hence, practically speaking, the computation will run for constant  $\alpha_k$ .

### Exercises

1. Show that among the  $\alpha_k, \gamma_k$  of the form  $\alpha_k = \alpha k^p, \gamma_k = \gamma k^r$  under the conditions of Theorem 1, the best choice, as to asymptotic convergence rate estimates, is the one given in the theorem:  $r = -1, p = -1/6$ .
2. Formulate an analog of Theorem 1 for a nonsymmetric difference approximation and under the assumption that  $\nabla f(x)$  satisfies a Lipschitz condition. Show that in this case the best choice of parameters is the following  $\gamma_k = \gamma/k, \alpha_k = \alpha/k^{1/4}$ , with  $E \|x^k - x^*\|^2 = O(1/k^{1/2})$ .

#### 4.4.3 Other Methods

For problems with noise all methods based on one-dimensional minimizations cease to be effective (for example, the methods of conjugate directions in Section 3.4) since such a minimization cannot be performed. There are more promising methods in which a nonlocal approximation of the function is constructed from its values at a number of points (such as the simplicial search method or the method of barycentric coordinates, see Section 3.4). The effect of the noise is that these methods cease to work in a neighborhood of the minimum where the noise level is comparable with the increments of the function. If the noise is random and centered, then the methods can be modified to remain efficient in that neighborhood. The general idea of the modification is to use a larger number of points in constructing the approximation of the function than in the deterministic case. This allows the noise to be averaged out and yields an ever more precise approximation. For example, in the simplicial method one can repeatedly calculate the function at each vertex of the simplex, comparing the accuracy of the estimated values of the function with their difference at distinct vertices.

A more economical method is to recompute the approximation after each new measurement. Let us just describe the scheme of such methods with a

simplified model. Suppose that it can be assumed that the function  $f(x)$ ,  $x \in \mathbf{R}^n$ , is affine in some region:  $f(x) \approx (a, x) + \beta$ , and that its values with noise have already been computed at  $k$  ( $k \geq n+1$ ) points:  $y_i = (a, x^i) + \beta + \eta_i$ ,  $i = 1, \dots, k$ , where  $\eta_i$  is random independent noise,  $E\eta_i = 0$ ,  $E\eta_i^2 = \sigma^2$ . Consider the  $(n+1)$ -dimensional vectors  $z^i = \{x^i, 1\}$ ,  $c^* = \{\alpha, \beta\}$  and write the measurements in the form  $y_i = (c^*, z^i) + \eta_i$ . We find a least squares estimate for  $c^*$ , i.e.,

$$\begin{aligned} c^k &= \underset{c}{\operatorname{argmin}} \sum_{i=1}^k (y_i - (c, z^i))^2 = \left( \sum_{i=1}^k z^i (z^i)^T \right)^{-1} \left( \sum_{i=1}^k z^i y_i \right) = \Gamma_k \sum_{i=1}^k z^i y_i, \\ \Gamma_k &= \left( \sum_{i=1}^k z^i (z^i)^T \right)^{-1}. \end{aligned} \quad (12)$$

This method can be given a recursive form—the new measurement at the point  $x^{k+1}$ :

$$y_{k+1} = (c^*, z^{k+1}) + \eta_{k+1}, \quad z^{k+1} = \{x^{k+1}, 1\},$$

can be taken into account by means of the formula

$$\begin{aligned} c^{k+1} &= c^k - \Gamma_{k+1} z^{k+1} ((c^k, z^{k+1}) - y_{k+1}), \\ \Gamma_{k+1} &= \Gamma_k - \frac{\Gamma_k z^{k+1} (\Gamma_k z^{k+1})^T}{1 + (\Gamma_k z^{k+1}, z^{k+1})}, \quad k \geq n+1, \\ \Gamma_{k+1} &= \left( \sum_{i=1}^{n+1} z^i (z^i)^T \right)^{-1}. \end{aligned} \quad (13)$$

Thus it is not necessary to solve the system of linear equations (12) to recompute the estimate at each step; rather it suffices to use the simple recurrence formula (13). The estimate  $c^k$  can be used to implement the step of descent:  $x^{k+1} = x^k - \gamma_k a^k$ ,  $c^k = \{a^k, \beta_k\}$ , and to verify the agreement of the linear model of the function with the measurements. Of course, in actual problems the linear model of the function is legitimate only locally, and the minimization method should include “forgetting” information obtained in earlier iterations.

Completely analogous methods can be applied to restoring a quadratic approximation of a function from measurements results containing random errors.

## 4.5 OPTIMAL METHODS IN THE PRESENCE OF NOISE

### 4.5.1 Potential Possibilities of Iterative Methods in the Presence of Noise

For deterministic “unperturbed” problems, as we have seen, there exists a set of methods each of which has its own intrinsic convergence rate. Thus for smooth strongly convex functions the heavy-ball method converges more rapidly than the gradient method, the conjugate gradient method more rapidly than the heavy ball method, Newton’s method more rapidly still, etc. The question of a convergence-rate optimal method is very complex. It turns out that the presence of noise in a certain sense simplifies the situation, inasmuch as it limits the possibilities of any of the minimization methods. In this case there exists a certain limiting convergence rate which cannot be surpassed. The method for which this limiting rate obtains is deemed optimal.

Let us begin with results establishing the *potential possibilities* for convergence rates of arbitrary iterative algorithms (not necessarily having to do with minimization) under random noises. Let us consider an iteration process in  $\mathbf{R}^n$ :

$$x^{k+1} = x^k - \gamma_k s^k, \quad s^k = R(x^k) + \xi^k, \quad (1)$$

where  $\gamma_k \geq 0$  are deterministic scalar factors,  $R(x)$  is some function, and  $\xi^k$  are random noises assumed to be independent and centered ( $E\xi^k = 0$ ). The initial approximation  $x^0$  can either be deterministic or random, and in the latter case it is assumed that  $E\|x^0\|^2 < \infty$  and  $x^0, \xi^i$  are independent. Suppose that there exists a unique point  $x^*$  such that  $R(x^*) = 0$  and  $R(x)$  satisfies the linear growth condition:

$$\|R(x)\| \leq L \|x - x^*\|. \quad (2)$$

**THEOREM 1.** For all  $k$  let

$$E\|\xi^k\|^2 \geq \sigma^2. \quad (3)$$

Then under the assumptions made above, for any method (1)

$$E\|x^k - x^*\|^2 \geq 1/(a + kb), \quad a = 1/E\|x^0 - x^*\|^2, \quad b = L^2/\sigma^2. \quad (4)$$

Note that in this theorem, in contrast with any theorems given previously, the convergence rate bounds are given from below, rather than from above. The theorem pertains to any way of *a priori* choosing  $\gamma_k$ , in particular to a choice where convergence fails to occur.

**PROOF.** Let us estimate the conditional mathematical expectation  $E(\|x^{k+1} - x^*\|^2 | x^k)$ :

$$\begin{aligned} E(\|x^{k+1} - x^*\|^2 | x^k) &= \|x^k - x^* - \gamma_k R(x^k)\|^2 + \gamma_k^2 E\|\xi^k\|^2, \\ \|x^k - x^* - \gamma_k R(x^k)\| &\geq (\|x^k - x^*\| - \gamma_k \|R(x^k)\|)_+, \\ &\geq (\|x^k - x^*\| - \gamma_k L \|x^k - x^*\|)_+, \\ E(\|x^{k+1} - x^*\|^2 | x^k) &\geq (1 - \gamma_k L)_+^2 \|x^k - x^*\|^2 + \gamma_k^2 \sigma^2. \end{aligned}$$

Then

$$E\|x^{k+1} - x^*\|^2 \geq (1 - \gamma_k L)_+^2 E\|x^k - x^*\|^2 + \gamma_k^2 \sigma^2.$$

The piecewise-quadratic function on the right attains a minimum with respect to  $\gamma_k$  for

$$\gamma_k^* = L E\|x^k - x^*\|^2 / (L^2 E\|x^k - x^*\|^2 + \sigma^2),$$

from which we obtain

$$\begin{aligned} E\|x^{k+1} - x^*\|^2 &\geq (1 - \gamma_k^* L)_+^2 E\|x^k - x^*\|^2 + (\gamma_k^*)^2 \sigma^2 \\ &= \sigma^2 E\|x^k - x^*\|^2 / (L^2 E\|x^k - x^*\|^2 + \sigma^2), \end{aligned}$$

or, denoting

$$u_k = 1/(E\|x^k - x^*\|^2), \quad u_{k+1} = L^2/\sigma^2 + u_k.$$

Thus,  $u_k \leq u_0 + kL^2/\sigma^2$ , i.e.,

$$E\|x^k - x^*\|^2 \geq [1/E\|x^0 - x^*\|^2 + kL^2/\sigma^2]^{-1}. \quad \square$$

From Theorem 1 it follows that any method of the form (1) cannot, under the conditions made above, converge faster than  $1/(a+bk)$ , or asymptotically faster than  $O(1/k)$ .

Let us give some examples of how this result is used. Again as in Section 4.2, we will consider the gradient method of minimizing  $f(x)$ :

$$x^{k+1} = x^k - \gamma_k s^k, \quad s^k = \nabla f(x^k) + \xi^k \quad (5)$$

under absolute random noise:

$$E\xi^k = 0, \quad E\|\xi^k\|^2 \geq \sigma^2 \quad (6)$$

(note the fact that here the inequality sign for the noise variance is reversed in comparison with Section 4.2). Suppose the  $f(x)$  has a minimum point  $x^*$  and the gradient  $\nabla f(x)$  satisfies the Lipschitz condition with constant  $L$ . Then the conditions for applicability of Theorem 1 obtain, and from it we deduce that for any choice of  $\gamma_k$ , for method (5) one has the estimate

$$\mathbf{E} \|x^k - x^*\|^2 \geq (1/\mathbf{E} \|x^0 - x^*\|^2 + kL^2/\sigma^2)^{-1}. \quad (7)$$

Differently put, no variant of the gradient method under absolute random noises can converge faster than  $O(1/k)$  (more precisely,  $\mathbf{E} \|x^k - x^*\|^2 \geq \sigma^2/(L^2 k) + o(1/k)$ ). Note that for the gradient method with  $\gamma_k = \gamma/k$  one had  $\mathbf{E} \|x^k - x^*\|^2 = O(1/k)$ , i.e., it is asymptotically optimal as regards convergence-rate order. The optimality of the gradient method will be investigated more accurately later.

Now let us consider Newton's method in the presence of noise. We assume that the matrix  $[\nabla^2 f(x^k)]^{-1}$  is computed precisely, and the gradient contains an additive random noise  $\xi^k$ . In this case Newton's method (modified by introduction of a parameter defining the step size) acquires the form

$$x^{k+1} = x^k - \gamma_k [\nabla^2 f(x^k)]^{-1} (\nabla f(x^k) + \xi^k). \quad (8)$$

We shall assume the noise  $\xi^k$  is independent and

$$\mathbf{E} \xi^k = 0, \quad \mathbf{E} \|\xi^k\|^2 \geq \sigma^2. \quad (9)$$

One can show that under the conditions of Theorem 1, Section 1.5, on convergence of the “unperturbed” Newton's method, the deterministic part of process (8), (i.e.,  $R(x^k) = [\nabla^2 f(x^k)]^{-1} \nabla f(x^k)$ ) satisfies a Lipschitz condition in a neighborhood of the solution, while the random part has variance bounded from below. Thus method (8) cannot converge faster than  $O(1/k)$ . Otherwise stated, the presence of random noises nullifies the advantages of rapidly convergent minimization processes.

Let us give a result analogous to Theorem 1 but for relative noise.

**THEOREM 2.** Let the assumptions formulated at the beginning of the section hold, and for all  $k$

$$\mathbf{E} \|\xi^k\|^2 \geq \tau \|x^k - x^*\|^2. \quad (10)$$

Then for any method (1),

$$\mathbf{E} \|x^k - x^*\|^2 \geq \mathbf{E} \|x^0 - x^*\|^2 q^k, \quad q = \tau/(L^2 + \tau). \quad \square \quad (11)$$

For the first example of application of Theorem 2 let us consider the gradient method in random relative noise. Let  $f(x)$  be differentiable, let the minimum point  $x^*$  exist, let  $\nabla f(x^k)$  satisfy a Lipschitz condition with constant  $L$ , and let the noise in determining the gradient be independent for different  $k$  and satisfy the conditions

$$\mathbf{E}\xi^k = \mathbf{0}, \quad \mathbf{E}\|\xi^k\|^2 \geq \tau \|x^k - x^*\|^2.$$

Then in method (5) inequality (11) holds for any  $\gamma_k$ . In other words, the gradient method under random relative noise cannot converge faster than the geometric progression.

Our second example will be the *random search method*. Let  $f(x)$  be a quadratic function:

$$f(x) = (Ax, x)/2 - (b, x), \quad \ell I \leq A \leq LI, \quad \ell > 0. \quad (12)$$

Consider the method

$$x^{k+1} = x^k - (\gamma_k/2(\alpha))(f(x^k) + \alpha h^k) - f(x^k - ah^k)h^k, \quad (13)$$

where  $h^k$  is a random vector uniformly distributed on the unit sphere, and  $\alpha > 0$  is the fixed size of the trial step. The method can be written in the form (see Section 3.4)

$$x^{k+1} = x^k - \gamma_k h^k (h^k)^T \nabla f(x^k) = x^k - \gamma_k s^k,$$

$$s^k = h^k (h^k)^T \nabla f(x^k).$$

Using the result of Exercise 1, we obtain

$$R(x^k) = \mathbf{E}s^k = \frac{1}{n} \nabla f(x^k),$$

$$\mathbf{E}\|\xi^k\|^2 = \mathbf{E}\|s^k - R(x^k)\|^2 = \frac{n-1}{n^2} \|\nabla f(x^k)\|^2 \geq \frac{n-1}{n^2} \ell^2 \|x^k - x^*\|^2.$$

From Theorem 2 it follows that, however  $\gamma_k$  may be chosen, the method of random search cannot converge more rapidly than the geometric progression with ratio

$$q = (n-1)\ell^2/(L^2 + (n-1)\ell^2). \quad (14)$$

In particular, for  $f(x) = \|x\|^2/2$ ,  $x \in \mathbf{R}^n$ , the method of random search cannot converge faster than the progression with ratio  $(n-1)/n$ .

Theorem 2 can be somewhat sharpened in case  $R(x)$  is linear, and a lower bound applies not only to the variance but also to the covariance matrix. Thus we consider the method

$$x^{k+1} = x^k - \Gamma_k(A(x^k - x^*) + \xi^k), \quad (15)$$

where  $\xi^k$  are independent,  $x^0$  is a random vector,  $A^{-1}$  exists and

$$\mathbf{E}\xi^k = 0, \quad \mathbf{E}\xi^k(\xi^k)^T \geq B > 0, \quad \mathbf{E}(x^0 - x^*)(x^0 - x^*)^T > 0, \quad (16)$$

and  $\Gamma_k$  are deterministic  $n \times n$  matrices.

**THEOREM 3.** In method (15) for any  $\Gamma_k$  one has the estimate

$$\begin{aligned} \mathbf{E}(x^k - x^*)(x^k - x^*)^T &\geq [(\mathbf{E}(x^0 - x^*)(x^0 - x^*)^T)^{-1} + kA^T B^{-1} A]^{-1} \\ &= \frac{1}{k} A^{-1} B (A^T)^{-1} + o\left(\frac{1}{k}\right). \quad \square \end{aligned} \quad (17)$$

As an application let us consider the generalization of the gradient method for minimizing the quadratic function

$$f(x) = (Ax, x)/2 - (b, x), \quad A \geq \ell I > 0,$$

under the noise:

$$x^{k+1} = x^k - \Gamma_k(\nabla f(x^k) + \xi^k), \quad \mathbf{E}\xi^k = 0, \quad \mathbf{E}\xi^k(\xi^k)^T = \sigma^2 I. \quad (18)$$

Applying Theorem 3, we obtain for any  $\Gamma_k$  that

$$\mathbf{E}(x^k - x^*)(x^k - x^*)^T \geq \left[ U_0^{-1} + \frac{k}{\sigma^2} A^2 \right]^{-1} = \frac{\sigma^2}{k} A^{-1} + o\left(\frac{1}{k}\right), \quad (19)$$

$$U_0 = \mathbf{E}(x^0 - x^*)(x^0 - x^*)^T,$$

$$\|\mathbf{E}(x^k - x^*)(x^k - x^*)^T\| \geq \frac{\sigma^2}{k\ell^2} + o\left(\frac{1}{k}\right), \quad (20)$$

where equality obtains in (19), (20) (see Exercise 2) for

$$\Gamma_k = (kA + \sigma^2 A^{-1} U_0^{-1})^{-1} = k^{-1} A^{-1} + o(1/k). \quad (21)$$

Comparing (20) with estimate (13) for the gradient method in Section 4.2, we find that under the present conditions the choice  $\gamma_k = 1/(k\ell)$  in the gradient method is asymptotically optimal.

## Exercises

- Let  $h$  be a random vector uniformly distributed on the unit sphere in  $\mathbf{R}^n$ . Show that  $Ehh^T = n^{-1}I$ , and if  $a$  is an arbitrary vector, then  $E\|hh^Ta - n^{-1}a\|^2 = (n^{-1} - n^{-2})\|a\|^2$ .
- Show that if  $E\xi^k(\xi^k)^T \neq B$ , then equality in (17) becomes equality for  $\Gamma_k$  defined by (21).

### 4.5.2 Optimal Algorithms

So far we have been limited to a very narrow class of algorithms: linear recursive algorithms. However, the problem of optimality can be resolved for much more general procedures. In a number of cases the potential minimization methods (not necessarily recursive or linear) with random noise can be established. The main tool here is the Cramer-Rao inequality known in statistics (the information inequality).

Let the function  $f(x)$  be quadratic

$$f(x) = (Ax, x)/2 - (b, x), \quad A > 0, \quad (22)$$

and its gradient be calculated with a random noise  $\xi$ . Suppose that the noise  $\xi$  is independent and identically distributed (earlier we made no such assumption). Suppose the values are already calculated as  $r^1 = \nabla f(x^1) + \xi^1$ , ...,  $r^k = \nabla f(x^k) + \xi^k$  at certain points  $x^1, \dots, x^k$ . Finally, let the matrices  $A$  and  $A^{-1}$  be known. Then  $x^i - x^* = A^{-1}r^i - A^{-1}\xi^i$ ,  $i = 1, \dots, k$ . Set  $z^i = x^i - A^{-1}r^i$ ,  $\eta^i = -A^{-1}\xi^i$ . Then  $z^i = x^* + \eta^i$ . The quantities  $z^i$  are known (since  $x^i$ ,  $r^i$  and  $A^{-1}$  are known), while the  $\eta^i$  are independent and identically distributed (since the  $\xi^i$  are). Thus the problem has been reduced to the following: Given the vectors  $z^i = x^* + \eta^i$ , where the  $\eta^i$  are realizations of an independent identically distributed random variable, it is required to estimate  $x^*$  from them.

This is the classic statistical problem of estimating parameters. The *Cramer-Rao inequality* is valid for it: If  $\eta^i$  have density  $p_\eta(z)$ , this density is regular (i.e., the equality  $\int \nabla p_\eta(z) dz = 0$  holds) and the *Fisher information matrix*  $J$  is nonsingular,

$$J = \int \frac{\nabla p_\eta(z) \nabla^T p_\eta(z)}{p_\eta(z)} dz, \quad 0 < J < \infty, \quad (23)$$

then for any unbiased estimate  $\hat{x}^k$  of the vector  $x^*$  from the measurements  $z^i$ ,  $i = 1, \dots, k$ , the inequality holds:

$$E(\hat{x}^k - x^*)(\hat{x}^k - x^*)^T \geq k^{-1}J^{-1}. \quad (24)$$

In other words, a lower bound to the accuracy of arbitrary unbiased estimates exists. Using (24) and the result of Problem 4, we arrive at the next result.

**THEOREM 4.** Let the noise  $\xi^i$  have density  $p(z)$ , where  $p(z)$  is regular and  $J = \int (\nabla p \nabla^T p)/p dz$  exists,  $0 < J < \infty$ . Then for any unbiased estimate  $\hat{x}^k$  of the minimum point  $x^*$  of the function (22) constructed from the measurements  $r^i = \nabla f(x^i) + \xi^i$ ,  $i = 1, \dots, k$ , at  $k$  points, the inequality holds:

$$E(\hat{x}^k - x^*)(\hat{x}^k - x^*)^T \geq k^{-1} A^{-1} J^{-1} A^{-1}. \quad \square \quad (25)$$

It is important that the measurement points  $x^1, \dots, x^k$  do not appear here. Thus no matter how the  $k$  points are chosen for measuring the gradient, the minimum cannot be found with accuracy greater than that given by inequality (25).

It remains to construct a method to obtain the indicated lower bound. Within the linear algorithms

$$x^{k+1} = \hat{x}_k^k \gamma_k H(\nabla f(x^k) + \xi^k), \quad (26) \quad \leftarrow x^k -$$

where  $H > 0$  is some matrix, then the asymptotically optimal choice of  $\gamma_k$  and  $H$  are the following:

$$\gamma_k = 1/k, \quad H = A^{-1}, \quad (27)$$

and here

$$E(x^k - x^*)(x^k - x^*)^T \leq k^{-1} A^{-1} B A^{-1} + o(k^{-1}), \quad B = E\xi\xi^T. \quad (28)$$

Noting Exercise 3, we obtain that if  $\xi^i$  are distributed normally, then the right side of (25) coincides with the right side of (28). Thus for the cases of normal noise (not just among linear or recursive algorithms), algorithm (26), (27) is asymptotically optimal. For other distributions of noise, algorithm (26), (27) is not generally optimal. Moreover, it can be shown that the right side of (25) is strictly less than the right side of (28) for any distribution other than normal. In this case an optimal algorithm can be obtained by introducing nonlinearity into the iterative process

$$x^{k+1} = x^k - \gamma_k \phi(\nabla f(x^k) + \xi^k), \quad (29)$$

where the function  $\phi: \mathbf{R}^n \rightarrow \mathbf{R}^n$  and  $\gamma_k$  are chosen as follows:

$$\phi(z) = -A^{-1} J^{-1} \nabla \log p(z), \quad \gamma_k = 1/k. \quad (30)$$

For normal noise method (29), (30) turns into (26), (27).

It can be shown that under certain conditions on  $p(z)$  the distribution of the variables  $\sqrt{k}(x^k - x^*)$  for method (29), (30) tends to normal distribution with zero mean and covariance matrix  $A^{-1}J^{-1}A^{-1}$ . Comparing this with the right side of (25), the method (29), (30) is seen to be asymptotically optimal.

Practical implementation of method (29), (30) is hampered by the fact that the matrix  $A^{-1}$  as well as the density of noise distribution must be known. We will not dwell on methods of overcoming these problems. Here it is more important that it is possible to construct an asymptotically optimal algorithm for solving a minimization problem under random noise, where the algorithm is recursive.

Let us further emphasize that all conclusions drawn here are asymptotic. The optimal algorithm for finite  $k$  in case of normal noise is given by expression (21). It is seen that at the early steps ( $k \ll \sigma^2 A^{-2} U_0^{-1}$ )  $\Gamma_k$  is roughly constant:  $\Gamma_k \approx \sigma^{-2} U_0 A$ , while for large  $k$ ,  $\Gamma_k$  decreases like  $k^{-1}$ :  $\Gamma_k = k^{-1} A^{-1} + o(k^{-1})$ .

Note also that the optimal algorithms presuppose exact knowledge of the distribution law of the noise and are unstable with respect to deviation of the true distribution from the supposed one. There are methods for overcoming this problem (the so-called *robust minimization algorithms*).

### Exercises

3. Let the random vector  $\eta$  be normally distributed with zero mean and covariance matrix  $S$ . Show that in this case the information matrix (23) is defined by the formula  $J = S^{-1}$ .
4. Let the random vectors  $\xi$  and  $\eta$  be related by the equality  $\eta = B\xi$ , where  $B$  is some matrix. Prove that for the corresponding information matrices one has  $J_\eta = BJ_\xi B^T$ .

$\checkmark$  nonsingular

## CHAPTER 5

### MINIMIZATION OF NONDIFFERENTIABLE FUNCTIONS

In many cases functions to be minimized turn out to be nondifferentiable. In later sections the reader will see examples of such functions, when we study decomposition, penalty functions, duality theorems, etc. Similarly, nonsmooth functions appear in the “best approximation” problems, parameter estimation problems by the least absolute-value criterion in statistics, in Steiner’s problem and related problems, among others. Frequently the objective function to be optimized (in engineering or economics) depends nondifferentiably on the parameters (for example, the dependence is often piecewise linear). Hence, in solving optimization problems one cannot restrict oneself to the case of smooth functions.

Undoubtedly, the problem of minimizing nondifferentiable functions in the general form is extremely complicated. These functions may be so ill-conditioned that their values on any finite set of points may not provide any information about the behavior of the function at other points. Therefore, we will deal here only with a special case of nonsmooth functions, viz. convex functions.

#### 5.1 CONVEX ANALYSIS: FUNDAMENTALS

Recently, mainly in the 1960s, a simple yet useful theory has been developed to work with convex functions usually referred to as *convex analysis*. We shall be using frequently the techniques of convex analysis. We begin with the basic notions of this theory.

### 5.1.1 Convex Sets and Projection

Recall that a set  $Q$  in  $\mathbf{R}^n$  is called convex if it contains any segment with the endpoints lying in  $Q$ , i.e., if for any

$$\begin{aligned} x, y &\in Q, \quad 0 \leq \lambda \leq 1, \\ \lambda x + (1 - \lambda)y &\in Q. \end{aligned} \tag{1}$$

By induction,  $Q$  contains also any convex combination of points, i.e.,

$$\begin{aligned} x^i &\in Q, \quad \sum_{i=1}^m \lambda_i = 1, \quad \lambda_i \geq 0, \\ i = 1, \dots, m \Rightarrow \sum_{i=1}^m \lambda_i x^i &\in Q. \end{aligned} \tag{2}$$

It is seen directly from the definition that a ball, a parallelepiped, a linear manifold, and a polyhedral set are convex, whereas the surface of a sphere or a finite collection of points are nonconvex (Fig. 10).

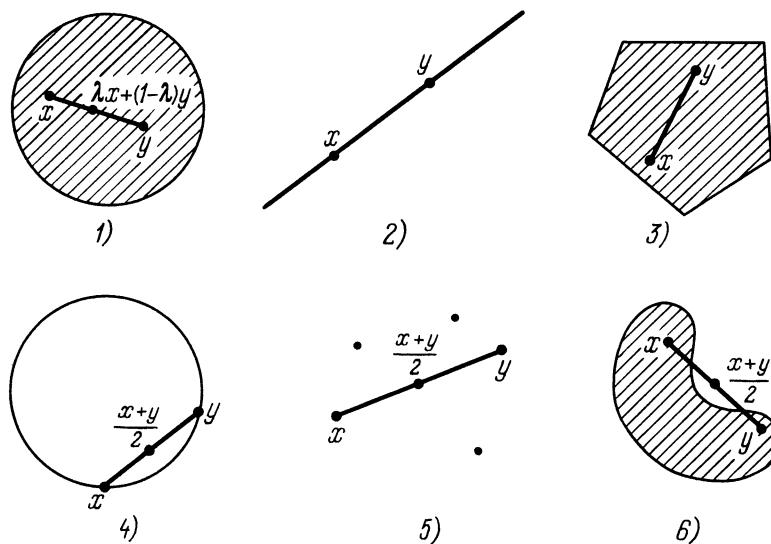


Fig. 10 Examples of convex (1-3) and nonconvex (4-6) sets.

For a convex function  $f(x)$  the set  $Q = \{x: f(x) \leq \alpha\}$  is obviously convex for any  $\alpha$ . The converse is not generally true: the function  $f(x) = \sqrt{\|x\|}$  is not convex but the sets  $\{x: f(x) \leq \alpha\}$  are convex (such functions are called *quasiconvex*).

If a set  $Q$  is nonconvex, it can be “convexified.” By the *convex hull*  $\text{Conv } Q$  of the set  $Q$  we mean the smallest convex set containing  $Q$ , i.e., the intersection of all convex sets containing  $Q$ . Such a set exists and is non-empty for nonempty  $Q$ . For instance, the convex hull of a sphere is a ball, the convex hull of two points is the segment joining them. It is not hard to verify that the convex hull can be defined differently, e.g., as the set of convex combinations of a finite number of points in  $Q$ , i.e.,

$$\text{Conv } Q = \left\{ x = \sum_{i=1}^m \lambda_i x^i; x^i \in Q, \lambda_i \geq 0, i = 1, \dots, m, \sum_{i=1}^m \lambda_i = 1 \right\}. \quad (3)$$

**LEMMA 1** (Caratheodory). For  $Q \subset \mathbf{R}^n$  in (3) one can take  $m \leq n+1$ .  $\square$

For a closed set  $Q$  the set  $\text{Conv } Q$  is not necessarily closed (for example, for

$$Q = \{x \in \mathbf{R}^2: x_2 = x_1^{1/2}, x_1 \geq 0\}$$

one has

$$\text{Conv } Q = \{x \in \mathbf{R}^2: 0 < x_2 = x_1^{1/2}, x_1 \geq 0\} \cup \{0, 0\} \quad \checkmark$$

$\checkmark$ )

**LEMMA 2.** If  $Q$  is closed and bounded, so is  $\text{Conv } Q$ .  $\square$

In what follows we will often use the projection operation. By the projection of the point  $x \in \mathbf{R}^n$  onto the set  $Q \subset \mathbf{R}^n$  we mean a point in  $Q$  (denoted  $P_Q(x)$ ) closest to  $x$ , i.e.,

$$P_Q(x) = \underset{y \in Q}{\operatorname{argmin}} \|x - y\|. \quad (4)$$

Clearly, if  $x \in Q$ , then  $P_Q(x) = x$ . Using the Weierstrass theorem (Section 1.3), we obtain that for closed  $Q$  the projection exists. If  $Q$  is convex, then the projection is unique since  $P_Q(x) = \underset{y \in Q}{\operatorname{argmin}} \phi(y)$ ,  $\phi(y) = \|x - y\|^2$  is a strictly convex function (Theorem 3 of Section 1.3). Finally, for a closed convex set  $Q$  the projection possesses the following properties (Fig. 11):

$$(x - P_Q(x), y - P_Q(x)) \leq 0 \quad \text{for all } y \in Q, \quad (5)$$

$$\|P_Q(x) - P_Q(y)\| \leq \|x - y\| \quad \text{for any } x, y. \quad (6)$$

## Exercises

- Prove that if  $Q$  is convex, then the sets  $\alpha Q = \{x = \alpha y, y \in Q\}$ ,  $AQ = \{x = Ay, y \in Q\}$  are convex (here  $\alpha \in \mathbb{R}^1$ ,  $A$  is an  $m \times n$  matrix), whereas if  $Q_1$  and  $Q_2$  are convex, then both  $Q_1 \cap Q_2$  and  $Q_1 + Q_2 = \{x = x_1 + x_2, x_1 \in Q_1, x_2 \in Q_2\}$  are convex.
- Prove that the function  $\rho_Q(x) = \|x - P_Q(x)\|$  is continuous for closed  $Q$  and convex for convex  $Q$ , whereas the function  $\phi(x) = \rho_Q^2(x)/2$  is convex and differentiable for closed convex  $Q$ , and  $\nabla \phi(x) = x - P_Q(x)$ .
- Let  $x$  be an interior point of the convex set  $Q$  and let  $y$  be a boundary point of  $Q$ . Prove that the points  $(1-\lambda)x + \lambda y$  are interior points of  $Q$  for  $0 \leq \lambda < 1$  and do not belong to  $Q$  for  $\lambda > 1$ .

### 5.1.2 Separation Theorems

Convex Analysis is based on the *separation theorems* (the *Hahn-Banach theorems*). Two sets  $Q_1$  and  $Q_2$  in  $\mathbb{R}^n$  are called *separable* if there is a hyperplane separating them (Fig. 12), or, in other words, if there is a number  $\alpha$  and a vector  $a \in \mathbb{R}^n$ ,  $a \neq 0$ , such that  $(a, x) \geq \alpha$  for all  $x \in Q_1$  and  $(a, x) \leq \alpha$  for all  $x \in Q_2$ . These sets are strictly separable if there are  $a \in \mathbb{R}^n$  and  $\alpha_1 > \alpha_2$  such that  $(a, x) \geq \alpha_1$  for  $x \in Q_1$  and  $(a, x) \leq \alpha_2$  for  $x \in Q_2$ .

**THEOREM 1** (separation theorem). Let  $Q_1, Q_2$  be convex disjoint sets in  $\mathbb{R}^n$  where  $Q_2$  is bounded. Then  $Q_1$  and  $Q_2$  are strictly separable.

**PROOF.** The function

$$\rho_1(x) = P_{Q_1}(x) = \|x - P_{Q_1}(x)\|,$$

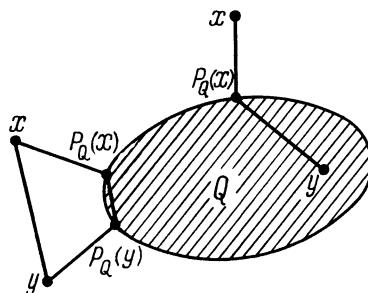


Fig. 11 Projections and its properties.

according to Exercise 2 above, is continuous. Hence, on the closed and bounded set  $Q_2$  it attains a minimum. Let

$$a_1 = P_{Q_1}(a_2), \quad a_2 = \underset{x \in Q_2}{\operatorname{argmin}} \rho_1(x).$$

Then  $a_1 \neq a_2$  (since  $Q_1$  and  $Q_2$  are disjoint),

$$\|a_1 - a_2\| = \rho(Q_1, Q_2) = \min \{\|x - y\|, x \in Q_1, y \in Q_2\}$$

and

$$a_2 = P_{Q_2}(a_1).$$

It follows from (5) that

$$(a_1 - a_2, x) \geq (a_1 - a_2, a_1) = \alpha_1 \quad \text{for } x \in Q_1,$$

$$(a_1 - a_2, x) \not\leq (a_1 - a_2, a_2) = \alpha_2 \quad \text{for } x \in Q_2,$$

$$\alpha_1 - \alpha_2 = \|a_1 - a_2\|^2 > 0.$$

Thus,  $Q_1$  and  $Q_2$  are strictly separable.  $\square$

This proof is completely graphic (see Fig. 12(b)). The boundedness condition on  $Q_2$  in the separation theorem cannot be dropped: the sets

$$Q_1 = \{x \in \mathbf{R}^2, x_2 \leq 0\}, \quad Q_2 = \{x \in \mathbf{R}^2, x_2 \geq x_1^{-1}, x_1 > 0\}$$

are not strictly separable.

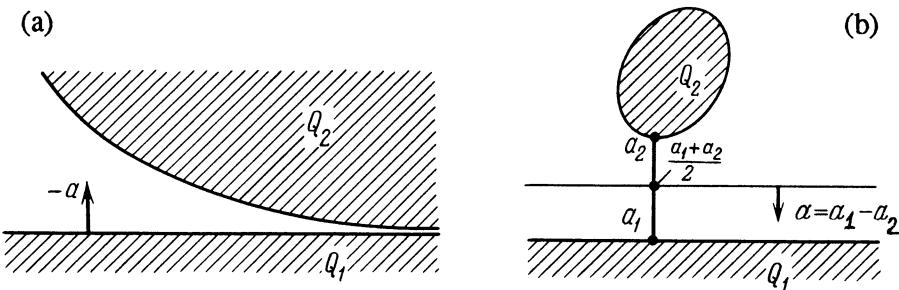


Fig. 12 Separation theorem: (a) separable sets;  
(b) strictly separable sets.

The separation theorem makes it possible to prove the next theorem on the supporting hyperplane. The hyperplane  $L = \{x: (a, x) = \alpha\}$  is called *supporting* for the set  $Q$  at the point  $x^0$  if  $x^0 \in L$  and all of the points of the set  $Q$  lie in the half-space defined by  $L$ , i.e.,  $(a, x) \leq \alpha$  for  $x \in Q$  (Fig. 13).

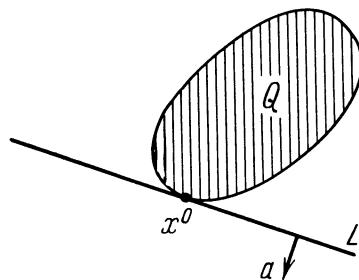


Fig. 13 The supporting hyperplane.

**THEOREM 2** (on the supporting hyperplane). Let  $Q$  be a convex set, and let  $x^0$  be a boundary point of  $Q$ . Then there exists a hyperplane supporting  $Q$  at  $x^0$ .  $\square$

### Exercises

4. Prove the following variants of the separation theorem:
  - (a) Let  $Q_1, Q_2$  be convex sets, and let  $Q_1$  and  $Q_2$  have interior points, none of which is common to either set. Then  $Q_1$  and  $Q_2$  are separable.
  - (b) Let  $Q_1, Q_2$  be convex disjoint sets. Then they are separable.
5. Show that the sets

$$Q_1 = \{x \in \mathbf{R}^2: |x_1| \leq 1, x_2 = 0\} \quad \text{and} \quad Q_2 = \{x \in \mathbf{R}^2: x_1 = 0, |x_2| \leq 1\}$$

are convex, have no common interior points, but are not separable (cf. Exercise 4(a)).

6. Prove that if  $x$  is a boundary point of  $\text{Conv } Q$ , then in Lemma 1, the  $n+1$  can be replaced by  $n$ .

### 5.1.3 Convex Nondifferentiable Functions

The definition of a convex function given in Section 1.1 also valid for nondifferentiable functions. Indeed, we say that a scalar function  $f(x)$

defined on the entire space  $\mathbf{R}^n$  is *convex* if for any  $x, y \in \mathbf{R}^n$  and  $0 < \lambda < 1$  we have the inequality

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \quad (7)$$

(see Fig. 1). Note that throughout this chapter we only consider functions whose domain of definition is the entire space. In Chapter 9 we shall introduce a more general class of convex functions defined on some set, for which many of the assertions of this chapter (for instance, Lemma 3) do not hold.

It is not hard to show that the class of convex functions is closed under the operations of addition, multiplication by a nonnegative number, and taking the maximum. Convex functions possess a number of other "advantageous" properties. In particular, it turns out that convex functions on  $\mathbf{R}^n$  are quite simple.

**LEMMA 3.** Any convex function on  $\mathbf{R}^n$  is continuous.

**PROOF.** Take arbitrary  $x \in \mathbf{R}^n$ ,  $\delta > 0$  and consider the points

$$a^1 = x + \delta e_1, \quad a^2 = x - \delta e_1, \quad \dots, \quad a^{2n-1} = x + \delta e_n, \quad a^{2n} = \delta e_n,$$

where  $e_1, \dots, e_n$  are the standard basis vectors. Let

$$\Delta(\delta) = \max_{1 \leq i \leq 2n} |f(a^i) - f(x)|.$$

Form the polyhedron  $Q(\delta)$  with vertices at these points:

$$Q(\delta) = \left\{ \sum_{i=1}^{2n} \mu_i a^i, \quad \mu_i \geq 0, \quad \sum_{i=1}^{2n} \mu_i = 1 \right\} = \left\{ x + \delta \sum_{i=1}^n \gamma_i e_i, \quad |\gamma_i| \leq 1 \right\}.$$


Let us prove that

$$\sup_{y \in Q(\delta)} |f(y) - f(x)| \leq \Delta(\delta).$$

Indeed, let

$$y = \sum_{i=1}^{2n} \mu_i a^i, \quad \mu_i \geq 0, \quad \sum_{i=1}^{2n} \mu_i = 1.$$

Then by Jensen's inequality (Lemma 1 of Section 1.1) one has

$$f(y) = \sum_{i=1}^{2n} \mu_i f(a^i) \leq \max_i f(a^i) \leq f(x) + \Delta(\delta).$$

On the other hand,  $f(y) \geq 2f(x) - f(y')$ , where  $y' \in Q(\delta)$  is the point symmetric to  $y$  with respect to  $x$ , i.e.,

$$\text{if } y = x + \delta \sum_{i=1}^n \gamma_i e_i, \quad \text{then } y' = x - \delta \sum_{i=1}^n \gamma_i e_i.$$

Hence

$$f(y) \geq 2f(x) - f(y') \geq f(x) - \Delta(\delta),$$

since  $f(y') \leq f(x) + \Delta(\delta)$  by what has been proved. Thus we do have  $|f(y) - f(x)| \leq \Delta(\delta)$  for all  $y \in Q(\delta)$ .

Next, we observe the following: any one-dimensional convex function  $\phi(\tau)$  is continuous. Indeed, for  $\varepsilon > 0$ ,

$$\begin{aligned} \phi(\tau + \varepsilon) &= \phi((1 - \varepsilon)\tau + \varepsilon(\tau + 1)) \leq (1 - \varepsilon)\phi(\tau) + \varepsilon\phi(\tau + 1) \\ &= \phi(\tau) + \varepsilon(\phi(\tau + 1) - \phi(\tau)); \end{aligned}$$

on the other hand,

$$\phi(\tau) = \phi\left[\frac{\varepsilon}{1+\varepsilon}(\tau-1) + \frac{1}{1+\varepsilon}(\tau+\varepsilon)\right] \leq \frac{\varepsilon}{1+\varepsilon}\phi(\tau-1) + \frac{1}{1+\varepsilon} \times \phi(\tau+\varepsilon),$$

i.e.,

$$\phi(\tau + \varepsilon) \geq \phi(\tau) + \varepsilon(\phi(\tau) - \phi(\tau-1)).$$

Hence  $\phi(\tau + \varepsilon) \rightarrow \phi(\tau)$  as  $\varepsilon \rightarrow +0$ . The case  $\varepsilon < 0$  is examined in the same way. Note that in this case the left and right derivatives  $\phi'_-(\tau) = \phi'(\tau; -1)$  and  $\phi'_+(\tau) = \phi'(\tau; +1)$  exist, and

$$\phi(\tau) - \phi(\tau - 1) \leq \phi'(\tau; 1) \leq \phi(\tau + 1) - \phi(\tau). \quad (8)$$

By continuity,

$$\Delta_i(\delta) = |f(a^i) - f(x)| = |f(x \pm \delta e_i) - f(x)|$$

tends to 0 as  $\delta \rightarrow 0$ . Hence also  $\Delta(\delta) = \max_i \Delta_i(\delta)$  tends to 0 as  $\delta \rightarrow 0$ . Therefore

$$\sup_{y \in Q(\delta)} |f(y) - f(x)| \rightarrow 0 \quad \text{as } \delta \rightarrow 0,$$

what was to be proved.  $\square$

**COROLLARY.** If  $f(x)$  is a convex function, then the set  $Q(\alpha) = \{x: f(x) \leq \alpha\}$  is convex and closed. In particular, the set  $X^* = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f(x)$  is convex and closed.  $\square$

**LEMMA 4.** The convex function  $f(x)$  at an arbitrary point  $x$  has a one-sided derivative in any direction uniformly bounded with respect to the directions:

$$\begin{aligned} f'(x; y) &= \lim_{\alpha \rightarrow +0} \frac{f(x+\alpha y) - f(x)}{\alpha} \leq f(x+y) - f(x) \\ &\leq \max_{\|z\|=\|y\|} (f(x+z) - f(x)). \quad \square \end{aligned} \tag{9}$$

### 5.1.4 The Subgradient

Of course, a convex function is not necessarily differentiable (Fig. 14). However, it is possible to use a notion similar in many respects to that of the gradient. Let  $f(x)$  be a function on  $\mathbb{R}^n$ . A vector  $a \in \mathbb{R}^n$  for which

$$f(x+y) \geq f(x) + (a, y) \tag{10}$$

for all  $y \in \mathbb{R}^n$  is called the *subgradient* of the function  $f(x)$  at the point  $x$ : we denote it  $\partial f(x)$ . As is seen in Figure 14, the subgradient is generally not defined uniquely. We will use  $\partial f(x)$  for the entire set of subgradients as well as for its arbitrary representation (one can usually see from the context which case it is). Let us now investigate properties of the subgradient.

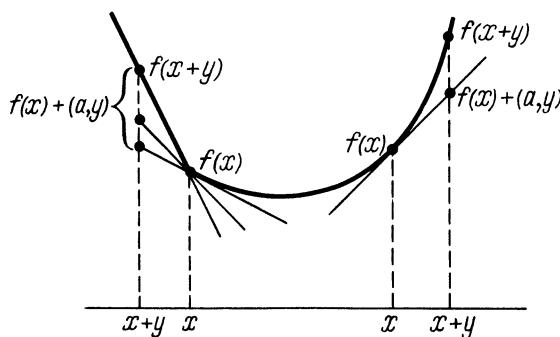


Fig. 14 The subgradient of a convex function.

**Lemma 5.** If  $f(x)$  is differentiable at a point  $x$ , then a subgradient is uniquely defined and coincides with the gradient:  $\partial f(x) = \nabla f(x)$ .

**PROOF.** Since the gradient satisfies inequality (26) in Section 1.1:  $f(x+y) \geq f(x) + (\nabla f(x), y)$ , it is a subgradient. Subtracting from inequality (10) the equality  $f(x+y) = f(x) + (\nabla f(x), y) + o(y)$  yields  $(\partial f(x) - \nabla f(x), y) \geq o(y)$ , which is possible for all  $y$  only if  $\partial f(x) - \nabla f(x) = 0$ .  $\square$

It can be shown that a convex function is differentiable almost everywhere (that is except on a set of measure zero). This is the well-known *Rademacher theorem*.

**LEMMA 6.** The set of subgradients at any point is nonempty, convex, closed and bounded.

Let us sketch the proof. Consider the set  $Q = \{x, \alpha: \alpha \geq f(x)\}$  in the space  $\mathbf{R}^{n+1}$  (this set is called the *epigraph* of the function  $f(x)$ ) (Fig. 15). The set  $Q$  is obviously convex, and Lemma 3 implies that it has interior points. The point  $\{x, f(x)\}$  is a boundary point of  $Q$ . By Theorem 2 there exists a supporting hyperplane for  $Q$  at this point, given by the vector  $\{a, -1\}$ . Thus  $a$  is a subgradient of  $f(x)$  at the point  $x$ . The convexity and closedness of the set of subgradients follows directly from the definition and the boundedness follows from Lemma 4.  $\square$

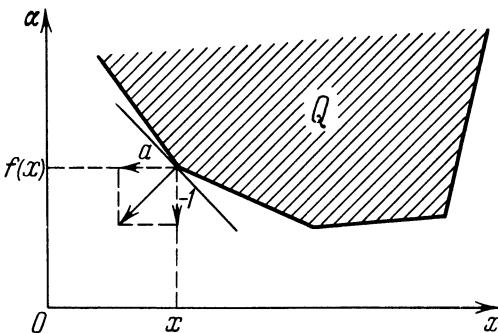


Fig. 15 Re: proof of the existence of the subgradient.

For nonsmooth functions we have the inequality similar to (29) in Section 1.1: for any  $x, y$

$$(\partial f(x) - \partial f(y), x - y) \geq 0, \quad (11)$$

i.e., the subgradient is a *monotone* operator.

When the subgradient is known one can compute the directional derivative (9) by a generalization of formula (6) in Section 1.1.

**LEMMA 7.** For any  $x, y$

$$f'(x; y) = \max_{a \in \partial f(x)} (a, y). \quad (12)$$

Next we sketch the proof of (12). Since  $f(x + \varepsilon y) - f(x) \geq \varepsilon(a, y)$  for all  $a \in \partial f(x)$ , then

$$f'(x; y) \geq \max_{a \in \partial f(x)} (a, y).$$

Assume there is a  $y^0$  such that

$$f'(x; y^0) > \max_{a \in \partial f(x)} (a, y^0).$$

Consider in the  $\mathbf{R}^{n+1}$  the ray

$$L = \{\alpha, z: \alpha = f(x) + \lambda f'(x; y^0), z = x + \lambda y^0, \lambda > 0\}$$

and the epigraph  $A = \{\alpha, z: \alpha > f(z)\}$ . Since  $f(z) \geq f(x) + \lambda f'(x; y^0)$ , the sets  $A$  and  $L$  are disjoint. Applying the separation theorem gives a contradiction.  $\square$

This and (6) in Section 1.1 yield the converse to Lemma 5: if  $\partial f(x)$  consists of one element, then  $f(x)$  is differentiable at  $x$ .

Lemmas 3, 4 and 7 lead to the following lemma.

**LEMMA 8.** The subgradients of a convex function  $f(x)$  are bounded on any bounded set or a set of the form  $\{x: f(x) \leq \underline{\alpha}\}$ .  $\square$

In what follows we will have to work sometimes with sums of sets (for example,  $\alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x)$  in Lemma 10 below). Recall that if  $A, B, C$  are sets in  $\mathbf{R}^n$ ,  $\beta, \gamma \in \mathbf{R}^1$ , then  $A = \beta B + \gamma C$  means that  $A = \{a = \beta b + \gamma c, b \in B, c \in C\}$ . We know (Exercise 1) that the sum of convex sets is convex;  $B+C = \emptyset$  if  $B = \emptyset$ .

**LEMMA 9.** If  $B$  and  $C$  are closed and bounded, then  $B+C$  is closed and bounded.  $\square$

The assumption of boundedness is essential in this case: e.g., if  $B = \{x \in \mathbf{R}^2: x_2 \geq x_1^{-1}, x_1 > 0\}$ ,  $C = \{x \in \mathbf{R}^2: x_1 = 0\}$ , then  $B$  and  $C$  are closed, but  $B+C = \{x \in \mathbf{R}^2: x_1 > 0\}$  is not closed.

Here are three lemmas that enable one to calculate subgradients of complex-valued functions.

 composite

 provided that  $f(x)$  is bounded from below

**LEMMA 10.** If  $f_1(x)$ ,  $f_2(x)$  are convex,  $f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x)$  and  $\alpha_1, \alpha_2 \geq 0$ , then

$$\partial f(x) = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x). \quad (13)$$

**PROOF.** The operation of directional differentiation is obviously linear:

$$f'(x; y) = \alpha_1 f'_1(x; y) + \alpha_2 f'_2(x; y) \quad \text{for all } x, y.$$

Next we use formula (12)

$$\begin{aligned} \max_{a \in \partial f(x)} (a, y) &= \max_{b \in \alpha_1 \partial f_1(x)} (b, y) + \max_{c \in \alpha_2 \partial f_2(x)} (c, y) \\ &= \max_{a \in \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x)} (a, y). \end{aligned}$$

By Lemmas 6, 9 and Exercise 1 the sets  $\partial f(x)$  and  $\alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x)$  are convex, closed and bounded. But if all  $y \in \mathbb{R}^n$ ,  $\max_{a \in A} (a, y) = \max_{b \in B} (b, y)$  for convex, closed and bounded sets  $A$  and  $B$ , then  $A$  and  $B$  are equal (this is easily proved, using Theorem 1). Hence (13) holds.  $\square$

Clearly, (13) extends to a sum of several convex functions:

$$\partial \left( \sum_{i=1}^m \alpha_i f_i(x) \right) = \sum_{i=1}^m \alpha_i \partial f_i(x), \quad \alpha_i \geq 0. \quad (14)$$

The next lemma gives a rule for computing the subgradient of the maximum of several functions.

**LEMMA 11.** Let

$$f(x) = \max_{1 \leq i \leq m} f_i(x),$$

where the  $f_i(x)$  are convex. Then

$$\partial f(x) = \text{Conv} \bigcup_{i \in I(x)} \partial f_i(x), \quad I(x) = \{i : f_i(x) = f(x)\}.$$

**PROOF.** By Lemmas 6, 2 the set

$$A = \text{Conv} \bigcup_{i \in I(x)} \partial f_i(x)$$

is convex, closed and bounded; so is the set  $\partial f(x)$ . It is not hard to see that

$$f'(x; y) = \max_{i \in I(x)} f'_i(x; y)$$

for all  $y$ . But by Lemma 7 and by the definition of  $\text{Conv } Q$ ,

$$\max_{i \in I(x)} f'_i(x; y) = \max_{\lambda_i \geq 0, \sum_i \lambda_i = 1} \sum_{i \in I(x)} \lambda_i f'_i(x; y) = \max_{a \in A} (a, y).$$

On the other hand,  $f'(x; y) = \max_{a \in \partial f(x)} (a, y)$  (Lemma 7). If

$$\max_{a \in A} (a, y) = \max_{a \in \partial f(x)} (a, y)$$

for all  $y$ , then (cf. Proof of Lemma 10)  $A = \partial f(x)$ .  $\square$

**LEMMA 12.** Let  $A$  be a  $m \times n$  matrix, let  $\phi(y)$  be a convex function on  $\mathbb{R}^m$ , and let  $f(x) = \phi(Ax)$ ,  $x \in \mathbb{R}^n$ . Then

$$\partial f(x) = A^T \partial \phi(Ax). \quad \square \quad (16)$$

Using Lemmas 10-12, the subgradients of varied functions can be calculated equally simply as the gradients of smooth functions can be calculated according to the usual rules of differentiation.

### Exercise

7. Calculate the subgradients of the following functions:

(a)  $f(x) = \|x\|$ ;

(b)  $f(x) = \sum_{i=1}^k |(a^i, x) - b_i|$ ;

(c)  $f(x) = \max_{1 \leq i \leq k} ((a^i, x) - b_i)$ .

L k

ANSWERS: (a)  $\partial f(x) = \begin{cases} \frac{x}{\|x\|}, & x \neq 0, \\ a, & \|a\| \leq 1, x = 0; \end{cases}$

(b)  $\partial f(x) = \sum_{i=1}^k \text{sign}((a^i, x) - b_i)a^i$ ;

(c)  $\partial f(x) = \sum_{i=1}^k \alpha_i a^i$ ,  $\alpha_i = 0$  for  $(a^i, x) - b_i < f(x)$ ,  $\alpha_i \geq 0$ ,  $\sum_{i=1}^k \alpha_i = 1$ .

### 5.1.5 The $\varepsilon$ -subgradient

The notion of a subgradient can be extended as follows. A vector  $a \in \mathbf{R}^n$  is called the  $\varepsilon$ -subgradient of the convex function  $f(x)$  at a point  $x$  if

$$f(x + y) \geq f(x) + (a, y) - \varepsilon \quad (17)$$

$\Rightarrow$  for all  $y \in \mathbf{R}^n$ . Here  $\varepsilon > 0$  is a fixed number. The set of  $\varepsilon$ -subgradients as well as an arbitrary  $\varepsilon$ -subgradient will be denoted  $\partial_\varepsilon f(x)$ . By definition,  $\partial f(x) = \partial_0 f(x)$ ,  $\partial f(x) \subset \partial_\varepsilon f(x)$  for all  $\varepsilon > 0$  and, furthermore,  $\partial f(x) = \bigcap_{\varepsilon > 0} \partial_\varepsilon f(x)$ . Graphically, the  $\varepsilon$ -subgradient corresponds to the hyperplanes in  $\mathbf{R}^{n+1}$  separating the epigraph of  $f(x)$  and the point  $\{f(x) - \varepsilon, x\}$  (Fig. 16). In contrast to the subgradient, the  $\varepsilon$ -subgradient for  $\varepsilon > 0$  is not determined by local properties of  $f(x)$ . Clearly, the  $\varepsilon$ -subgradient is not unique even for differentiable functions; the affine function  $f(x) = (c, x) + \alpha$  is the exception. Here  $\partial_\varepsilon f(x) \equiv c$  for all  $\varepsilon, x$ .

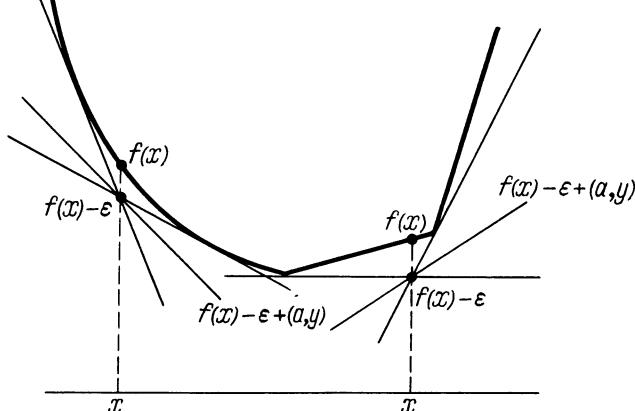
Rules for calculating the  $\varepsilon$ -subgradients are not as simple as for subgradients. We give here one important particular case where finding a  $\varepsilon$ -subgradient requires less calculations than the subgradient. Let

$$f(x) = \max_{y \in Q} \phi(x, y), \quad (18)$$

where  $x \in \mathbf{R}^n$ ,  $Q$  is a compact set,  $\phi(x, y)$  is continuous in  $y$  and convex in  $x$ . In particular,  $Q$  may consist of a finite number of elements (then we obtain the function given in Lemma 11). Obviously,  $f(x)$  is defined on  $\mathbf{R}^n$  and is convex. Let  $y = y(x)$  be any point in  $Q$  such that

$$\phi(x, y) \geq f(x) - \varepsilon. \quad (19)$$

Fig. 16 The  $\varepsilon$ -subgradient.



In other words,  $\bar{y}$  is an arbitrary point at which the maximum of  $\phi(x, y)$  in  $\bar{y} \in Q$  is attained approximately (to within  $\varepsilon$ ).

**LEMMA 13.**

$$\partial_x^V(x, \bar{y}) \subset \partial_\varepsilon f(x). \quad (20) \quad \checkmark \phi$$

**PROOF.** For any  $z$ , by the definition of the subgradient and by (18), (19) we have

$$\begin{aligned} f(x+z) &= \max_{y \in Q} \phi(x+z, y) \geq \phi(x+z, \bar{y}) \geq \phi(x, \bar{y}) + (\partial_x \phi(x, \bar{y}), z) \\ &\geq f(x) + (\partial_x \phi(x, \bar{y}), z) - \varepsilon. \quad \square \end{aligned}$$

Thus, to find one of the  $\varepsilon$ -subgradients of  $f(x)$  of the form (18), it suffices to find approximately the maximum in  $y$  and take the subgradient of the respective function  $\phi$ , whereas the calculation of the subgradient of  $f(x)$  requires that  $\phi$  be maximized exactly in  $y$ .

## 5.2 EXTREMUM CONDITIONS, EXISTENCE, UNIQUENESS, AND STABILITY OF A SOLUTION

To analyze the problem

$$\min f(x), \quad x \in \mathbf{R}^n, \quad (1)$$

where  $f(x)$  is a convex nondifferentiable function on  $\mathbf{R}^n$ , we follow the lines of Sections 1.2 and 1.3 for smooth functions.

### 5.2.1 Extremum Conditions

It is easy to formulate necessary and sufficient conditions for the minimum in terms of subgradients.

**THEOREM 1.** The condition

$$0 \in \partial f(x^*) \quad (2)$$

is necessary and sufficient for the point  $x^*$  to be a solution of (1).

**PROOF.** N e c e s s i t y. Let  $x^*$  be a minimum point of  $f(x)$ . Then  $f(x^* + y) \geq f(x^*) + (0, y)$  for all  $y$ . This means ((10) in Section 5.1) that 0 is the subgradient of  $f(x)$  at  $x^*$ .

S u f f i c i e n c y. If 0 is the subgradient at  $x^*$ , then  $f(x^* + y) \geq f(x^*) + (0, y) = f(x^*)$  for all  $y$ , i.e.,  $x^*$  is a solution of (1).  $\square$

L0

Of course, there may also be nonzero subgradients at a minimum point (e.g., for  $f(x) = \|x\|$ :  $\partial f(x) = \{a: \|a\| \leq 1\}$ , see Exercise 7 of Section 5.1), and this is the difference between condition (2) and the condition  $\nabla f(x) = 0$  for smooth functions. In other words, extremum conditions in the nonsmooth case do not reduce to the solution of equations. Thus the assertion stated in Section 1.3 becomes even more lucid: extremum conditions are not designed for finding a minimum.

Using the notion of a  $\varepsilon$ -subgradient, we can formulate the necessary and sufficient conditions for the point  $x_\varepsilon$  to be an approximate solution of problem (1).

**THEOREM 2.** The condition

$$0 \in \partial_\varepsilon f(x_\varepsilon) \quad (3)$$

holds iff

$$f(x_\varepsilon) \leq \inf_x f(x) + \varepsilon. \quad \square$$

### Exercises

1. Check the validity of the following extremum conditions:

(a)  $f(x) = \sum_{i=1}^m \alpha_i \|x - a^i\|$ ,  $\alpha_i > 0$ ,  $x, a^i \in \mathbb{R}^n$ . Then  $\nabla f(x^*) = 0$ , if  $x^* \neq a^i$ , and  $\alpha_i \geq \|\sum_{j \neq i} (\nabla_j(a^i - a^j)) / \|a^i - a^j\|$ , if  $x^* = a^i$ .

(b)  $f(x) = \sum_{i=1}^m |(a^i, x) - b_i|$ . There are  $|\lambda_i^*| \leq 1$ ,  $i \in I^* = \{i: (a^i, x^*) = b_i\}$  such that  $\sum_{i \in I^*} \lambda_i^* a^i + \sum_{i \in I_+} a^i - \sum_{i \in I_-} a^i = 0$ ,  $I_+ = \{i: (a^i, x^*) \neq b_i\}$ ,  $I_- = \{i: (a^i, x^*) \neq b_i\}$ .

(c)  $f(x) = \max_{1 \leq i \leq m} f_i(x)$ , where  $f_i(x)$  are convex differentiable functions.

Then there exist  $\lambda_i^* \geq 0$ ,  $i \in I^* = \{i: f_i(x^*) = f(x^*)\}$ ,  $\sum_{i \in I^*} \lambda_i^* = 1$  such that  $\sum_{i \in I^*} \lambda_i^* \nabla f_i(x^*) = 0$ .

2. Let  $f(x)$  be the same as in Exercise 1(c). For the point  $x_\varepsilon$  let

$$\lambda_i \geq 0, \quad i \in I_\varepsilon = \{i: f_i(x_\varepsilon) \geq f(x_\varepsilon) - \varepsilon\}, \quad \sum_{i \in I_\varepsilon} \lambda_i = 1$$

such that  $\sum_{i \in I_\varepsilon} \lambda_i \nabla f_i(x_\varepsilon) = 0$ . Then  $f(x_\varepsilon) \leq \inf_x f(x) + \varepsilon$ . To prove, use Lemma 13 of Section 5.1 and Theorem 2.

### 5.2.2 Existence and Uniqueness of a Minimum

**THEOREM 3.** Let the function  $f(x)$  be convex on  $\mathbf{R}^n$ , and let the set  $Q_\alpha = \{x: f(x) \leq \alpha\}$  nonempty and bounded for some  $\alpha$ . Then  $f(x)$  attains a minimum on  $\mathbf{R}^n$ .

Indeed, by Lemma 3 of Section 5.1,  $f(x)$  is continuous, and therefore Weierstrass' theorem is applicable (Section 1.3).  $\square$

It is easy to solve the problem on the uniqueness of a minimum for strictly convex functions. Recall (Section 1.1) that a function is strictly convex if for any  $x \neq y$ ,  $0 < \lambda < 1$ ,

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y). \quad (4)$$

**THEOREM 4.** A minimum point of a strictly convex function is unique. The proof is obvious.  $\square$

### Exercises

3. Prove that for a strictly convex function the following inequality holds for all  $y \neq 0$ :

$$f(x+y) > f(x) + (\partial f(x), y). \quad (5)$$

4. Show that the function  $f(x) = \|x\|$  is not strictly convex, whereas the function  $\sum_{i=1}^m \alpha_i \|x-a^i\|$ ,  $\alpha_i > 0$ , is strictly convex provided the points  $a^i$  are not collinear.

5. Prove that the function

$$f(x) = \sum_{i=1}^m \alpha_i \|x-a^i\|, \quad \alpha_i > 0,$$

attains a minimum on  $\mathbf{R}^n$  and is unique if the  $a^i$  are not collinear.

### 5.2.3 Stability of a Minimum

**THEOREM 5.** The unique minimum point of a convex function is globally stable, i.e., any minimizing sequence converges to it. The bounded set of minimum points  $X^*$  is weakly stable, i.e., any minimizing sequence has limit points which belong to  $X^*$ .  $\square$

These assertions follow directly from the continuity of  $f(x)$  (Lemma 3 of Section 5.1) and the following easily verifiable fact.

**LEMMA 1.** If  $Q_\alpha = \{x: f(x) \leq \alpha\}$  is bounded and nonempty for some  $\alpha$  for the convex function  $f(x)$ , then  $Q_\alpha$  is bounded for all  $\alpha$ .  $\square$

Quantitative estimates of stability are easily obtainable for a class of strongly convex functions. Recall the definition of strong convexity given in Section 1.1 relating to smooth and also nonsmooth functions  $f(x)$ : we can find an  $\ell > 0$  such that

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) - \ell\lambda(1-\lambda)\|x - y\|^2/2 \quad (6)$$

for any  $x, y$  and  $0 \leq \lambda \leq 1$ . Such functions have the following properties.

**LEMMA 2.** For a strongly convex function  $f(x)$  we have

$$f(y) \geq f(x) + (\partial f(x), y - x) + \ell\|y - x\|^2/2 \quad (7)$$

for all  $x, y$ , the  $f(x)$  attaining a unique minimum  $x^*$  and for all  $x$

$$f(x) \geq f(x^*) + \ell\|x - x^*\|^2/2. \quad (8)$$

**PROOF.** By the definition of a subgradient we have

$$f(\lambda x + (1-\lambda)y) = f(x + (1-\lambda)(y - x)) \geq f(x) + (1-\lambda)(\partial f(x), y - x).$$

$\angle 6$  Substituting this inequality into (7) and cancelling out the term  $1 - \lambda$  yield

$$f(y) \geq f(x) + (\partial f(x), y - x) + \ell\lambda\|x - y\|^2/2.$$

This holds for all  $\lambda < 1$ ; passing to the limit as  $\lambda \rightarrow 1$ , we obtain (7). It follows from (7) that  $Q = \{y: f(y) \leq f(x)\}$  is bounded, and Theorems 3 and 4 imply the existence and uniqueness of  $x^*$ . Using Theorem 1 together with (7), we arrive at (8).  $\square$

Inequality (8) makes it possible to estimate the proximity of  $x$  to  $x^*$  from that of  $f(x)$  to  $f(x^*)$ . A particular case of (8) for smooth functions was given in Section 1.3.

It is however worth noting that for nonsmooth problems the strong convexity property is in general not typical. There is another important class of functions for which stability can be guaranteed; this class includes nonsmooth functions only. We say that  $x^*$  is a *sharp minimum point* of  $f(x)$  if for all  $x$  (Fig. 17)

$$f(x) \geq f(x^*) + \alpha\|x - x^*\|, \quad \alpha > 0. \quad (9)$$

This condition cannot be satisfied *a priori* for smooth functions (Exercise 8 in Section 1.3).

**LEMMA 3.** The following conditions are equivalent to (9) for a convex function  $f(x)$ :

- (a)  $f'(x^*; y) \geq \alpha > 0$  for all  $y$ ;
- (b) 0 is an interior point of  $\partial f(x^*)$ .  $\square$

Using (9), one can estimate the proximity of  $x$  to  $x^*$ , knowing how close  $f(x)$  is to  $f(x^*)$ . However the *superstability* property of a sharp minimum is of greater interest; this property does not hold for the problems involving strongly convex functions. A sharp minimum point is invariant under small perturbation of the function.

**THEOREM 6.** Let  $f(x)$  be a convex function on  $\mathbb{R}^n$ , let  $x^*$  be a sharp minimum point, and let  $g(x)$  be a convex function. Then we can find an  $\varepsilon_0 > 0$  such that for  $0 \leq \varepsilon < \varepsilon_0$  the minimum point of the function  $f(x) + \varepsilon g(x)$  is unique and coincides with  $x^*$ .

**PROOF.** By Lemma 10 of Section 5.1, for  $\phi_\varepsilon(x) = f(x) + \varepsilon g(x)$  we have  $\partial\phi_\varepsilon(x) = \partial f(x) + \varepsilon\partial g(x)$ . Since 0 is an interior point of  $\partial f(x^*)$  (Lemma 3), while  $\partial g(x^*)$  is bounded (Lemma 6 of Section 5.1), then for sufficiently small  $\varepsilon$  one has  $\varepsilon\partial g(x^*) \subset -\partial f(x^*)$ , i.e.,  $0 \in \partial\phi_\varepsilon(x^*)$ . By Theorem 1,  $x^*$  is a minimum point of  $\phi_\varepsilon(x)$ .  $\square$

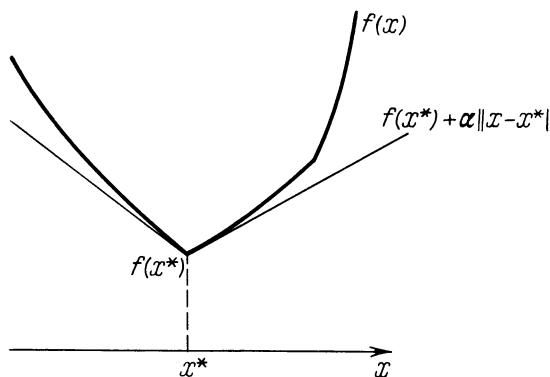


Fig. 17 The sharp minimum.

## Exercises

6. Prove the following generalization of Theorem 6. Let  $f(x)$  be a convex function and let  $X^* = \operatorname{Argm}_{\substack{x \in \mathbb{R}^n}} f(x) \neq \emptyset$ . Furthermore, let

$$f(x) \geq f^* + \alpha \rho(x, X^*) , \quad \alpha > 0 ,$$

where  $f^* = f(x^*)$ ,  $x^* \in X^*$ ,  $\rho(x, X^*) = \|x - P_{X^*}(x)\|$ . Also, let  $g(x)$  be a convex function and let the  $X_g^* = \operatorname{Argmin}_{\substack{x \in X^*}} g(x)$  be nonempty and bounded. Then

$$X_g^* = \operatorname{Argm}_{\substack{x \in \mathbb{R}^n}} [f(x) + \varepsilon g(x)]$$

for sufficiently small  $\varepsilon > 0$ .

7. Analyze the notion of a condition number of a minimum point (Section 1.3) for nonsmooth  $f(x)$ . What does  $\mu$  equal to for  $f(x) = \sum_{i=1}^n \lambda_i |x_i|$ ,  $\lambda_i > 0$ ?

## 5.3 THE SUBGRADIENT METHOD

### 5.3.1 The Substance of the Method

The fundamental algorithms for minimizing smooth functions, the gradient as well as Newton's algorithms, are based on linear or quadratic approximation of the function given by the first terms of a Taylor series. However this method is unfeasible for nondifferentiable functions, for such a function cannot be well approximated either by a linear or by a quadratic function. The methods for minimizing smooth functions, described in Chapter 3, become ineffective when one passes to nondifferentiable functions. Here are a few examples.

Let  $f(x) = |x_1 - x_2| + 0.2|x_1 + x_2|$  be a function of two variables. Then at the point  $\{1, 1\}$  its values along either coordinate axis increase, but this point is not a minimum point (Fig. 18). Hence the method of coordinatewise descent is inapplicable to minimizing nondifferentiable functions.

One might try to construct an analog of the steepest descent method. The vector  $s = s(x) \in \mathbb{R}^n$ ,  $\|s\| = 1$ , is called the *direction of steepest descent* at the point  $x$  if this is indeed the direction in which the functional  $f(x)$  decreases most rapidly:

$$s(x) = \operatorname{argmin}_{\|y\|=1} f'(x; y) . \quad (1)$$

By formula (12) of Section 5.1, for a convex function the direction of steepest descent exists and is defined by

$$s = -P_{\partial f(x)}(0)/\|P_{\partial f(x)}(0)\| , \quad (2)$$

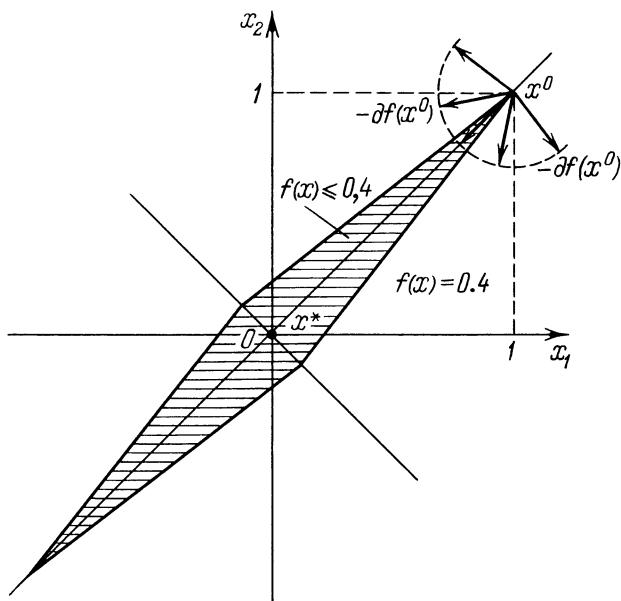


Fig. 18 Difficulties in minimizing a nonsmooth function.

i.e.,  $s$  is the subgradient with minimal norm. However, it is possible to construct a convex function for which the steepest descent method

$$x^{k+1} = x^k + \gamma_k s^k(x^k), \quad \gamma_k = \underset{\gamma > 0}{\operatorname{argmin}} f(x^k + \gamma s(x^k))$$

"jams" without reaching the minimum point.

Methods for minimizing nonsmooth functions cannot be further developed without new, innovative techniques. N.Z. Shor suggests—however surprisingly—a direct analog of the gradient method, with the gradient replaced by an arbitrary subgradient of the function:

$$x^{k+1} = x^k - \gamma_k \partial f(x^k). \quad (3)$$

We consider again the function  $f(x) = |x_1 - x_2| + 0.2|x_1 + x_2|$ . Then the vector  $\{1.2; -0.8\}$  is a subgradient at the point  $\{1; 1\}$ ; however, the motion along the subgradient makes the function increase for any choice of the step size  $\gamma_k$  (Fig. 18). Thus, the values of the function in method (3) cannot decrease monotonically. In this case, however, another function, viz. the distance to the minimum point, decreases monotonically. This is the key idea of the subgradient method (3). The rule for choosing the step size is also of special interest. It is clear that in (3)  $\gamma_k \equiv \gamma$  is not

possible, in contrast to the gradient method. For example, for the function  $f(x) = \|x\|$  we have  $\|\partial f(x)\| = 1$  for all  $x \neq 0$ , and therefore  $\|x^{k+1} - x^k\| \equiv \gamma$ ; hence the method does not converge. On the other hand, it is impossible to choose  $\gamma_k$  to be the same as in the steepest descent method, for the  $f(x)$  does not necessarily decrease in the direction  $-\partial f(x^k)$ . In the subgradient method it is possible to reduce the step size either by using the proximity of the value of the function at the current point to the minimum, or by choosing some sequence *a priori* tending to 0. We shall examine both methods in what follows.

### 5.3.2 The Main Results

Let  $f(x)$  be a convex function. Also, let us assume that some subgradient  $\partial f(x^k)$  can be computed at a point  $x^k$ . We consider the *subgradient method* in the following form:

$$x^{k+1} = x^k - \gamma_k \frac{\partial f(x^k)}{\|\partial f(x^k)\|}, \quad \gamma_k \rightarrow 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty. \quad (4)$$

In other words, the step of fixed size  $\gamma_k$  is made from the point  $x^k$  in the opposite direction from the subgradient. The step size tends to 0, whereas the total step size is infinite. Examples of the sequences  $\gamma_k$  satisfying conditions (4) are given by

$$\gamma_k = \frac{\gamma}{k+c}, \quad \gamma_k = \frac{\gamma}{k^\rho}, \quad 0 < \rho \leq 1, \quad \gamma_k = \frac{\gamma}{k \ln k}. \quad (5)$$

The assertion on convergence in Theorem 1 (and in many cases in the sequel) concerns the quantity

$$\phi_k = \min_{0 \leq i \leq k} f(x^i) \quad (6)$$

being the *record value* of  $f(x)$  over  $k$  iterations.

**THEOREM 1.** In method (4) for convex  $f(x)$ :  $\phi_k \rightarrow f^* = \inf_{x \in \mathbb{R}^n} f(x)$ .

We emphasize the fact that in this case there is no need for either existence of the minimum point or, *a fortiori*, lower boundedness of  $f(x)$ ; it is possible that  $f^* = -\infty$ .

**PROOF.** Suppose that  $f(x^k) \geq \tilde{f}$  for all  $k$  and some  $\tilde{f} > f^*$ . Take the point  $\tilde{x}$  such that  $f(\tilde{x}) < \tilde{f}$ . By the continuity of  $f(x)$  (Lemma 3 in Section 5.1) we can find a  $\rho > 0$  such that  $f(x) \leq \tilde{f}$  for  $\|x - \tilde{x}\| \leq \rho$ . In particular, for

$x_\rho = \tilde{x} + \rho \partial f(x^k) / \|\partial f(x^k)\|$  we have  $f(x_\rho) \leq \tilde{f}$ . On the other hand,

$$\begin{aligned} f(x_\rho) &\geq f(x^k) + (\partial f(x^k), x_\rho - x^k) \geq \\ &\geq \tilde{f} + (\partial f(x^k), \tilde{x} - x^k) + (\partial f(x^k), x_\rho - \tilde{x}) \\ &= \tilde{f} + (\partial f(x^k), \tilde{x} - x^k) + \rho \|\partial f(x^k)\|, \end{aligned}$$
✓

i.e.,

$$(\partial f(x^k), x^k - \tilde{x}) / \|\partial f(x^k)\| \geq \rho.$$

Let us now estimate the distance to  $\tilde{x}$  in the iterations:

$$\begin{aligned} \|x^{k+1} - \tilde{x}\|^2 &= \|x^k - \tilde{x}\|^2 - 2\gamma_k \left[ \frac{\partial f(x^k)}{\|\partial f(x^k)\|}, x^k - \tilde{x} \right] + \gamma_k^2 \\ &\leq \|x^k - \tilde{x}\|^2 - 2\gamma_k \rho + \gamma_k^2. \end{aligned}$$

Since  $\gamma_k \rightarrow 0$ , we can find a  $k_0$  such that  $\gamma_k \leq \rho$  for  $k \geq k_0$ . Hence for  $k \geq k_0$ , we have

$$\|x^{k+1} - \tilde{x}\|^2 \leq \|x^k - \tilde{x}\|^2 - \gamma_k \rho.$$

Summing these inequalities over  $k$ , we obtain  $\rho \sum_{k=k_0}^{\infty} \gamma_k \leq \|x^{k_0} - \tilde{x}\|^2$ , which contradicts the condition  $\sum_{k=0}^{\infty} \gamma_k = \infty$ . Thus, the inequality  $f(x^k) \geq \tilde{f} > f^*$  is impossible for all  $k$ , which is equivalent to the condition  $\phi_k \rightarrow f^*$ . □

One can also derive convergence assertions for  $x^k$  for the nonempty set of minimum points  $X^*$  (Exercise 1 below).

Clearly, method (4) cannot converge rapidly—in other words, the distance to the minimum point cannot be less than the step size  $\gamma_k$ ; this quantity decreases slowly since the condition  $\sum_{k=0}^{\infty} \gamma_k = \infty$  must be satisfied. In particular, it is possible to show that in method (4) there is *a priori* no convergence with the rate of geometric progression. Furthermore, the choice of  $\gamma_k$  from the conditions  $\gamma_k \rightarrow 0$ ,  $\sum_{k=0}^{\infty} \gamma_k = \infty$  is inappropriate, because there are many similar sequences and it is not quite clear which one to choose. Hence we will describe other possible methods for adjusting the step size.

In some problems the minimal value of the function may be known (we denote it by  $f^*$ ). Thus, for instance, if the system of compatible linear equations

$$(a^i, x) = b_i, \quad i = 1, \dots, n, \quad x \in \mathbf{R}^n,$$

is reduced to a minimization of the function

$$f(x) = \sum_{i=1}^n |(a^i, x) - b_i|$$

or of the function

$$f(x) = \max_{1 \leq i \leq n} |(a^i, x) - b_i|,$$

then  $f^* = 0$  in both cases. The value  $f^*$  makes it possible to construct the following variant of the subgradient method that contains no arbitrary parameters:

$$x^{k+1} = x^k - \frac{f(x^k) - f^*}{\|\partial f(x^k)\|^2} \partial f(x^k). \quad (7)$$

This choice of the step size is shown graphically in Figure 19.

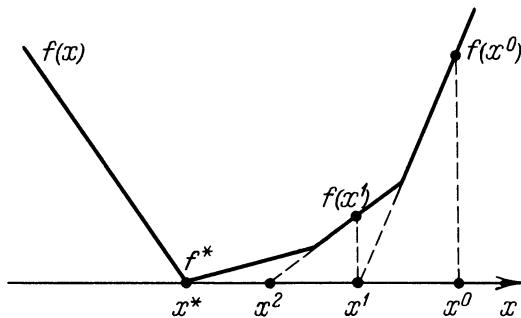


Fig. 19 The method for choosing a step size in the subgradient method.

**THEOREM 2.** Let  $f(x)$  be a convex function on  $\mathbf{R}^n$ , whose set of minimum points  $X^*$  is nonempty. Then in method (7),  $x^k \rightarrow x^* \in X^*$ . The estimation of the convergence rate is as follows: for an arbitrary function  $f$ ,

$$\lim_{k \rightarrow \infty} \sqrt{k} (f(x^k) - f^*) = 0; \quad (8)$$

for the function with a sharp minimum one can claim convergence with the rate of geometric progression.

**PROOF.** Let  $\tilde{x}$  be an arbitrary minimum point. Then

$$\begin{aligned} \|x^{k+1} - \tilde{x}\|^2 &= \|x^k - \tilde{x}\|^2 - 2 \frac{(\partial f(x^k), x^k - \tilde{x})(f(x^k) - f^*)}{\|\partial f(x^k)\|^2} \\ &\quad + \frac{(f(x^k) - f^*)^2}{\|\partial f(x^k)\|^2} \\ &\leq \|x^k - \tilde{x}\|^2 - \frac{(f(x^k) - f^*)^2}{\|\partial f(x^k)\|^2} \end{aligned} \tag{9}$$

and  $(f(x^k) - f^*)/\|\partial f(x^k)\| \rightarrow 0$ . Since the sequence  $x^k$  is bounded:  $\|x^k - \tilde{x}\| \leq \|x^0 - \tilde{x}\|$ , then (Lemma 8 of Section 5.1)  $\|\partial f(x^k)\| \leq c$ . Hence  $f(x^k) \rightarrow f^*$ . Therefore, we can find a sequence  $x^{k_i} \rightarrow x^*$ , where  $x^*$  is a minimum point. If in the foregoing estimate we replace  $\tilde{x}$  by  $x^*$ ,  $\|x^k - x^*\|$  will monotonically decrease, whereas  $\|x^{k_i} - x^*\| \rightarrow 0$ . Thus  $x^k \rightarrow x^*$ .

We proceed to estimate the rate of convergence. From (9) we have

$$\sum_{k=0}^{\infty} \frac{(f(x^k) - f^*)^2}{\|\partial f(x^k)\|^2} < \infty,$$

and from the boundedness of  $\|\partial f(x^k)\|$  we have  $\sum_{k=0}^{\infty} (f(x^k) - f^*)^2 < \infty$ . If we assume that  $\lim_{k \rightarrow \infty} \sqrt{k} (f(x^k) - f^*) > 0$ , then  $f(x^k) - f^* > a/\sqrt{k}$  for sufficiently large  $k$ , which contradicts the condition  $\sum_{k=0}^{\infty} (f(x^k) - f^*)^2 < \infty$ . Thus,  $\lim_{k \rightarrow \infty} \sqrt{k} (f(x^k) - f^*) = 0$ .

Next, let  $f(x)$  have a sharp minimum, i.e.,  $f(x) - f^* \geq \alpha \|x - x^*\|$ . Then

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - (\alpha^2/c^2) \|x^k - x^*\|^2 = q \|x^k - x^*\|^2,$$

$$q = 1 - \alpha^2/c^2,$$

which proves the convergence with the rate of geometric progression.  $\square$

The ratio of this progression may, however, be very close to 1 if the level lines of  $f(x)$  are strongly elongated (i.e., if the minimization problem is ill-posed).

When  $f^*$  is unknown, the method may be modified; for example, it is possible to apply the iterative process

$$x^{k+1} = x^k - \frac{f(x^k) - \bar{f}}{\|\partial f(x^k)\|^2} \partial f(x^k), \tag{10}$$

where  $\bar{f}$  is some estimate of  $f^*$ , and  $\bar{f}$  is updated on the basis of the behavior of the  $x^k$ .

As was noted earlier, the iterative process (7) can be applied similarly to minimizing smooth convex functions, and its rate of convergence is of the same order for other “good” variants of the gradient method (see (34) of Section 3.3).

### Exercises

1. Prove the following variants of Theorem 1 ( $f(x)$  is assumed to be convex and  $X^*$  nonempty):

- (a) If  $X^*$  is bounded, then  $\rho(x^k, X^*) \rightarrow 0$ .
- (b) If  $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$ , then  $x^k \rightarrow x^* \in X^*$ .
- (c) If  $(X^*) \neq \emptyset$ , then the method is finite.

*Hint:* (a) use Lemma 6' of Section 2.2; (b) use Lemma 2 of Section 2.2.

2. What can be said of the behavior of the following methods?

- (a)  $x^{k+1} = x^k - \gamma \partial f(x^k) / \|\partial f(x^k)\|$ ,  $\gamma > 0$ ;
- (b)  $x^{k+1} = x^k - \gamma_k \partial f(x^k)$ ,  $\gamma_k \rightarrow 0$ ,  $\sum_{k=0}^{\infty} \gamma_k = \infty$ ;
- (c)  $x^{k+1} = x^k - \gamma_k \partial f(x^k) / \|\partial f(x^k)\|$ ,  $\gamma_k = \gamma_0 q^k$ ,  $q < 1$ ;
- (d)  $x^{k+1} = x^k - \gamma(f(x^k) - f^*) \partial f(x^k) / \|\partial f(x^k)\|^2$ .

ANSWERS: (a) The method “converges to within  $\gamma$ ,” i.e., there exists a function  $\psi(\gamma) > 0$ ,  $\psi(\gamma) \rightarrow 0$  as  $\gamma \rightarrow 0$  such that

$$\lim_{k \rightarrow \infty} \phi_k \leq f^* + \psi(\gamma), \quad \phi_k = \min_{1 \leq i \leq k} f(x^i).$$

- (b) If  $\|\partial f(x)\| \leq c$  for all  $x$ , then Theorem 1 holds.

(c) For the case of a sharp minimum, for a given  $x^0$  one can choose  $\gamma_0$  and  $q$  such that the method converges with the rate of geometric progression.

- (d) For  $0 < \gamma < 2$  Theorem 2 holds.

3. Prove that if  $f(x)$  is convex and  $X^*$  is nonempty and bounded and  $\gamma_k = \gamma/\sqrt{k}$ , then in method (4)  $\phi_k - f^* = O(1/\sqrt{k})$ . *Hint:* For the quantity  $\phi_{km} = \min_{m \leq i \leq k} (f(x^i) - f^*)$ ,  $m < k$ , obtain the bound

$$\phi_{km} = c(\|x^m - x^*\|^2 + \sum_{i=m}^{\infty} \gamma_i^2) / \sum_{i=m}^k \gamma_i$$

and choose  $m = k/2$  for even  $k$ .

#### 5.3.3 The $\varepsilon$ -subgradient Method

Examine whether it is possible to replace the subgradient by the  $\varepsilon$ -subgradient in methods of the form (3). It would be appropriate to do

so because in many problems it is easier to calculate the  $\varepsilon$ -subgradient than the gradient (see Lemma 13 of Section 5.1).

The most straightforward approach involves a substitution of an arbitrary  $\varepsilon$ -subgradient for the  $\partial f(x)$  in method (4). However, if  $\varepsilon$  is fixed, then the new method may not converge: e.g., by Theorem 2 of Section 5.2 the  $\varepsilon$ -subgradient may vanish at any point at which the value of  $f(x)$  differs from the optimum by less than  $\varepsilon$ —the method can therefore stop at such points. To make this method converge, we need to vary  $\varepsilon$  by letting it go to 0 in the iterative process. Then we obtain the following method:

$$\begin{aligned} x^{k+1} &= x^k - \gamma_k \frac{\partial_{\varepsilon_k} f(x^k)}{\|\partial_{\varepsilon_k} f(x^k)\|}, \\ \gamma_k &\rightarrow 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty, \quad \varepsilon_k \rightarrow 0 \quad \text{as } k \rightarrow \infty. \end{aligned} \tag{11}$$

**THEOREM 3.** In method (11) for the convex function  $f(x)$  we have

$$\phi_k = \min_{1 \leq i \leq k} f(x^i) \rightarrow \inf f(x).$$

Proof follows that of Theorem 1.  $\square$

## 5.4 ALTERNATIVE METHODS

### 5.4.1 Preliminary Remarks

As was shown before, the subgradient method is very simple and it converges under weak assumptions concerning the function. However its rate of convergence may be poor. Note first that for smooth functions the subgradient method turns into the gradient method, the only difference from the standard variants of the latter being in the rules for choosing the step size. As we have seen earlier, the gradient method is ineffective for ill-conditioned functions. Secondly, the subgradient method in the form (4) in Section 5.3 cannot converge rapidly (even at the rate of geometric progression) for any function. Moreover, the variant (7) of the subgradient method, as the proof of Theorem 2 implies, converges slowly, too (as the geometric progression with ratio close to 1), for ill-conditioned nonsmooth functions. Thus, the subgradient method cannot be an effective tool for solving convex nondifferentiable problems. More powerful optimization methods are in order.

In the smooth case, these methods have been patterned after Newton's method, i.e., based on a quadratic approximation of the objective function. In the case of nonsmooth problems one has to take a different approach, for

example, a piecewise linear approximation that is normal for nondifferentiable functions. But the set of nonsmooth convex functions has too much variety to be well approximated by the class of piecewise linear functions. This limits the capability of this approach. The problem of minimizing an arbitrary convex function is, in general, too involved. Hence the method that uses only subgradients and rapidly converges for all functions of a given class is theoretically unfeasible.

### 5.4.2 Multistep Methods

The simplest technique for improving convergence is to exploit the information obtained in the preceding iterations. Assume that the points  $x^0, \dots, x^k$  have been constructed and the subgradients  $\partial f(x^0), \dots, \partial f(x^k)$  have been computed. Using the relations

$$f(x^*) \geq f(x^i) + (\partial f(x^i), x^* - x^i),$$

we can assert that the minimum point  $x^*$  lies in the region defined by the linear inequalities

$$Q_k = \{x: (\partial f(x^i), x - x^i) \leq f^* - f(x^i), i = 0, \dots, k\}, \quad (1)$$

and for the unknown  $f^* = f(x^*)$  it lies in the broader domain

$$Q = \{x: (\partial f(x^i), x - x^i) \leq 0, i = 0, \dots, k\}. \quad (2)$$

In order to reduce this region to its minimum (Fig. 20), a new point  $x^{k+1}$  can be added. This can be done in many ways. In what follows we shall describe variants of these methods and give results concerning their convergence, to demonstrate the convergence of the quantity  $\phi_k - f^*$  to 0 with a specified rate, where

$$f^* = \min_{x \in \mathbb{R}^n} f(x), \quad \phi_k = \min_{0 \leq i \leq k} f(x^i).$$

In all these methods the polyhedron  $Q_0$  is assumed to contain  $x^*$ , which is the region of *a priori* localization of the minimum. In the implementation, it is usually easy to identify the possible range of variation of each variable; let this parallelepiped be  $Q_0$ .

In the *cutting-plane method* the point  $x^{k+1}$  is the minimum point of the piecewise linear approximation of  $f(x)$  defined by the values of  $f(x^i)$  and of  $\partial f(x^i)$ ,  $i = 0, \dots, k$ , on the set  $Q_0$ . In other words,  $x^{k+1}$  is the solution of the linear programming problem:

$$\min z,$$

$$f(x^i) + (\partial f(x^i), x - x^i) \leq z, \quad i = 0, \dots, k, \quad x \in Q_0. \quad (3)$$

Here  $z \in \mathbf{R}^1$  is an auxiliary variable equal to the ordinate of the approximating function (Fig. 21). In this method, in contrast to other methods, we have to solve an auxiliary linear programming problem in each iteration; this means a problem of constrained minimization. This is indeed typical of nonsmooth problems which require piecewise linear approximation. We shall see later that in constrained problems it is common to use methods based on a reduction of the problem at hand to a unconstrained minimization problem. This is not a contradiction because the resulting auxiliary problems are simpler than the initial problems. To estimate the efficiency of the methods in this case, one has to evaluate accurately how difficult it is to solve the auxiliary problem.

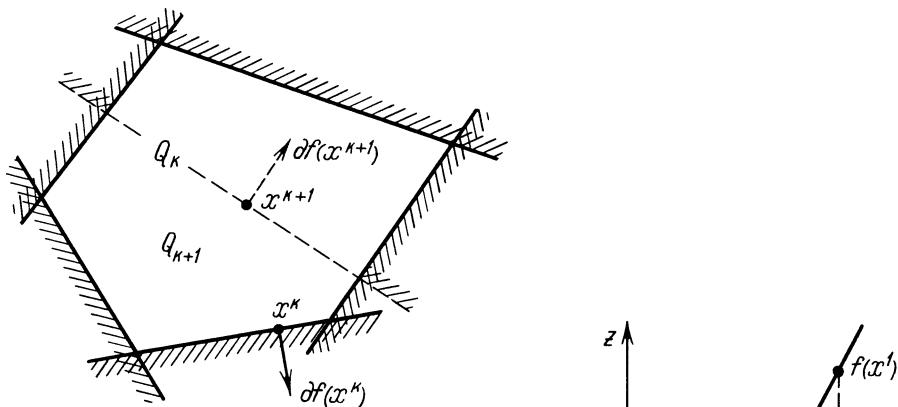


Fig. 20 A general scheme of the cutoff method.

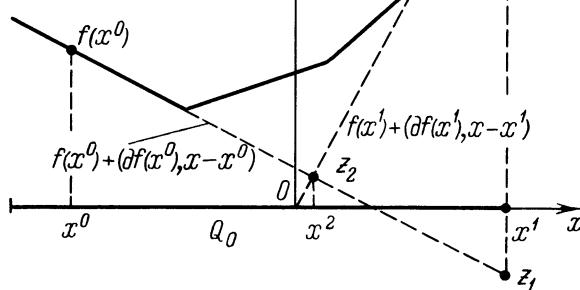


Fig. 21 The cutting hyperplane method.

**THEOREM 1.** Let  $f(x)$  be a convex function on  $\mathbf{R}^n$  and let the set  $Q_0$  be bounded and contain a minimum point  $x^*$ . Then in method (3) we have  $\phi_k \rightarrow f^*$ .

**PROOF.** Let  $z_{k+1}, x^{k+1}$  be the solution of problem (3). Then

$$z_{k+1} \leq \max_{0 \leq i \leq k} [f(x^i) + (\partial f(x^i), x - x^i)]$$

for all  $x \in Q_0$  (Fig. 21) and, in particular,

$$z_{k+1} \leq \max_{0 \leq i \leq k} [f(x^i) + (\partial f(x^i), x^* - x^i)].$$

By the convexity of  $f(x)$  we have

$$f(x^*) \geq f(x^i) + (\partial f(x^i), x^* - x^i), \quad 0 \leq i \leq k,$$

i.e.,

$$f^* \geq \max_{0 \leq i \leq k} [f(x^i) + (\partial f(x^i), x^* - x^i)].$$

A comparison of these inequalities yields  $z_{k+1} \leq f^*$ . On the other hand,  $f(x^{k+1}) \geq f^*$ , i.e.,  $z_{k+1} \leq f^* \leq f(x^{k+1})$ . Suppose that  $f(x^{k+1}) - z_{k+1} \geq \varepsilon > 0$  for all  $k \geq k_0$ . Then

$$\begin{aligned} f(x^i) &\geq f(x^{k+1}) + (\partial f(x^{k+1}), x^i - x^{k+1}) \\ &\geq z_{k+1} + \varepsilon + (\partial f(x^{k+1}), x^i - x^{k+1}) \\ &\geq f(x^i) + (\partial f(x^i), x^{k+1} - x^i) + \varepsilon + (\partial f(x^{k+1}), x^i - x^{k+1}) \\ &\geq f(x^i) + \varepsilon - 2L\|x^i - x^{k+1}\|, \end{aligned}$$

where

$$L = \max_{x \in Q_0} \|\partial f(x)\|$$

(this quantity is bounded by Lemma 8 of Section 5.1). Hence  $\|x^{k+1} - x^i\| \geq \varepsilon/2L$  for all  $i = 0, \dots, k$  and all  $k \geq k_0$ . This contradicts the compactness of  $Q_0$ . Hence

$$\lim_{k \rightarrow \infty} (f(x^k) - z_k) = 0,$$

and since

$$0 \leq f(x^k) - f^* \leq f(x^k) - z_k ,$$

then

$$\lim_{k \rightarrow \infty} f(x^k) = f^* . \quad \square$$

The question of rate of convergence of this method has been given so far little attention. For some problems (say, problems with a sharp minimum) the method obviously, converges rapidly. For piecewise linear problems it is finite. However in the general case the convergence rate is very small. Consider the one-dimensional problem

$$\min p^{-1}x^p , \quad 0 \leq x \leq 1, \quad x_0 = 0, \quad x_1 = 1 .$$

Each auxiliary problem of (3) has a nonunique solution, the  $x_{k+1}$  being the largest solution. Then  $x_{k+1} = x_k - p^{-1}x_k = qx_k$ ,  $x_k = q^{k-1}x_1$ ,  $q = 1-p^{-1}$ , and for large  $p$  the progression ratio  $q$  is close to 1. For multidimensional problems the linear convergence rate cannot apparently be guaranteed even for smooth strongly convex functions.

The drawback of this method is the need to solve linear programming problems with an increasing number of constraints. One may modify the method so as to remove this drawback: roughly, keep only those constraints which can be satisfied as equalities. Or, in solving a subsequent problem use the solution of the preceding problem as the initial approximation—to do this go to the dual problem.

An alternative method for choosing the point  $x^{k+1}$  is employed in the *method of Chebyshev centers*, in which a polyhedron  $Q_k$  of the form (1) or (2) is taken as the Chebyshev center, i.e., the point the maximum distance from which to the faces of the polyhedron is minimal. In other words,  $x^{k+1}$  is the solution of the problem

$$\max z ,$$

$$\left[ \frac{\partial f(x^i)}{\|\partial f(x^i)\|}, x - x^* \right] + z \leq 0 , \quad i = 0, \dots, k, \quad x \in Q_0 , \quad (4)$$

or, if  $f^*$  is known, of the problem

$$\max z ,$$

$$\left[ \frac{\partial f(x^i)}{\|\partial f(x^i)\|}, x - x^* \right] + z \leq f^* - f(x^i) , \quad i = 0, \dots, k, \quad x \in Q_0 . \quad (5)$$

It is possible to show that for (4) and (5) an analog of Theorem 1 holds true. Regarding the convergence rate of the method we can easily see that in the one-dimensional case method (4) becomes the dichotomy method:  $x^{k+1}$  is taken as the midpoint of the minimal segment with endpoints  $x^i$  for which the  $\partial f(x^i)$  has different signs at these endpoints (Fig. 22(a)). This implies that method (4), unlike method (3), is not finite for piecewise linear  $f(x)$ . For the multidimensional case, as is seen in Figure 22(b), the convergence is slow.

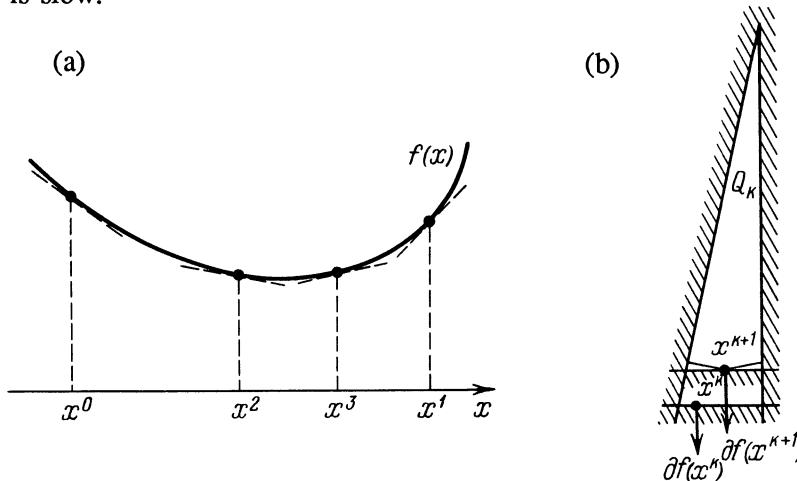


Fig. 22 The method of Chebyshev centers:  
(a) one-dimensional case; (b) two-dimensional case.

It is also possible to construct the point  $x^{k+1}$  in a different way. Let us take any set of indices  $I$  from the set  $0, \dots, k$ , for instance,  $I = \{0, \dots, k\}$ , or  $I = \{k\}$  or  $I = \{k, k-1\}$ . As  $x^{k+1}$  we take the point closest to  $x^k$  and satisfying constraints of the form (1) for the set  $I$ . In other words,  $x^{k+1}$  is the solution of the problem

$$\begin{aligned} \min & \|x - x^k\|^2, \\ f(x^i) + (\partial f(x^i), x - x^i) & \leq f^*, \quad i \in I \end{aligned} \quad (6)$$

(assume that  $f^*$  is known). The auxiliary problem (6) is a quadratic programming problem with an objective function of the form  $\|x - a\|^2$ . Thus, it is reduced to the projection of  $x^k$  onto the polyhedron given by the linear constraints (6). It is convenient to go from this problem over to the dual problem, viz. find the solution to be (see Section 10.4)

$$x^{k+1} = x^k - \sum_{i \in I} \lambda_i^k \partial f(x^i), \quad (7)$$

where  $\lambda_i^k$  is the solution of the problem

$$\min_{\substack{\lambda_i \geq 0 \\ i \in I}} \left[ \left\| \sum_{i \in I} \lambda_i \partial f(x^i) \right\|^2 - 4 \sum_{i \in I} \lambda_i (\partial f(x^i), x^k - x^i) - 4 \sum_{i \notin I} \lambda_i (f(x^i) - f^*) \right]. \quad (8)$$

To solve this problem of minimizing the quadratic function on the nonnegative orthant is quite simple (see Section 7.3). Clearly, if  $I = \{k\}$ , then the method coincides with the subgradient method (7) of Section 5.2.

Method (6) is superior to methods (3), (4), (5) since  $I$  need not contain all of the preceding indices, and the auxiliary problems to be solved at each step can be of small dimension. However, the need to know  $f^*$  is the disadvantage of this method.

An ingenious technique of choosing  $x^{k+1}$  is used in the *center-of-gravity* method. Let

$$Q_k = \{x \in Q_0 : (\partial f(x^i), x - x^i) \leq 0, i = 1, \dots, k\}, \quad (9)$$

$x^{k+1}$  being the center of gravity of  $Q_k$ .

The choice is due to the following result in the theory of convex bodies.

**LEMMA 1.** Let  $Q$  be a convex body (i.e., a set with nonempty interior) in  $\mathbb{R}^n$ ,  $a$  being the center of gravity,  $L$  being the hyperplane passing through  $a$ ,  $v_1$  and  $v_2$  being the subsets into which  $L$  divides  $Q$  (Fig. 23). Then

$$\frac{v_i}{v} \leq 1 - \left( \frac{n}{1+n} \right)^n < 1 - \frac{1}{e}, \quad i = 1, 2, \quad v = v_1 + v_2. \quad (10)$$

For points other than  $a$  the right side in (10) can only be greater.  $\square$

In other words, the volume of the subset “truncated” from  $Q$  by the hyperplane passing through the center of gravity is never smaller than the  $e^{-1}$  of  $Q$ , and may be smaller for the remaining points. This is exactly the reason for the choice of  $x^{k+1}$  as the center of gravity of  $Q_k$ .

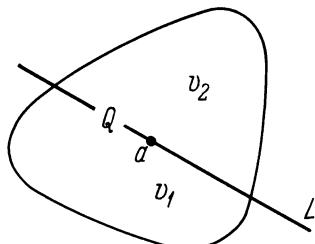


Fig. 23 The lemma on the center of gravity.

**THEOREM 2.** Let  $f(x)$  be a convex function on  $\mathbf{R}^n$  and let  $Q_0$  be a bounded, closed and convex set. Then in method (9) one has

$$\begin{aligned}\phi_k - f^* &\leq cq^k, \\ q &= \left(1 - \left(\frac{n}{n+1}\right)^n\right)^{1/n} < \left(1 - \frac{1}{e}\right)^{1/n} = 1 - \frac{1}{ne} + o\left(\frac{1}{n}\right), \\ c &= \max_{x \in Q_0} (f(x) - f^*). \end{aligned}\tag{11}$$

**PROOF.** By Lemma 1, the volume  $v_k$  of the polyhedron  $Q_k$  satisfies the inequality  $v_{k+1} \leq v_k \beta$ ,  $\beta = 1 - (n/(n+1))^n$ , i.e.,  $v_k \leq v_0 \beta^k$ . Take an arbitrary minimum point  $x^* \in Q_k$  and construct the set  $S$  from  $Q_k$  by a similarity transformation with center at  $x^*$  and extension coefficient  $\alpha = \beta^{-k/n}$ , i.e.,  $S = \{x: x^* + \alpha y, x^* + y \in Q_k\}$ . Then its volume  $v(S) = \alpha^n v_k \leq \alpha^n v_0 \beta^k = v_0$ . Hence the set  $Q_0$  cannot fit strictly the set  $S$  and therefore there is a  $z \in Q_0$ ,  $z \notin S^0$ . This implies that  $u = (1 - \alpha^{-1})x^* + \alpha^{-1}z \notin Q_k^0$  (since  $z$  is obtained from  $u$  by the extension). But if  $u \notin Q_k^0$ , then (by the definition of  $Q_k$ ) we can find an  $i$ ,  $1 \leq i \leq k$ , such that  $(\partial f(x^i), u - x^i) \geq 0$ . Hence

$$f(u) \geq f(x^i) + (\partial f(x^i), u - x^i) \geq f(x^i) \geq \phi_k.$$

Using the convexity of  $f(x)$ , we obtain

$$\begin{aligned}\phi_k &\leq f(u) = f((1 - \alpha^{-1})x^* + \alpha^{-1}z) \\ &\leq (1 - \alpha^{-1})f^* + \alpha^{-1}f(z) \leq f^* + c/\alpha, \\ c &= \max_{x \in Q_0} (f(x) - f^*), \end{aligned}$$

where  $c < \infty$  since the  $f(x)$  is continuous and the  $Q_0$  is bounded. Thus

$$\phi_k - f^* \leq c\alpha^{-1} = c\left(1 - \left(\frac{n}{n+1}\right)^n\right)^{k/n}. \quad \square$$

For  $n = 1$  the set  $Q_k$  is a segment and  $x^{k+1}$  is its midpoint. Hence the center-of-gravity method becomes the dichotomy method. For  $n = 2$  a method for finding the center of gravity would be based on the fact that the center of gravity of a triangle is given by an intersection of its meridians, while the center of gravity of two joint configurations is found by the formula  $\tilde{x} = \alpha \tilde{x}_1 + (1 - \alpha) \tilde{x}_2$ , where  $\tilde{x}$ ,  $\tilde{x}_1$ ,  $\tilde{x}_2$  are the centers of gravity of  $A$ ,  $A_1$ ,  $A_2$  (with  $A = A_1 \cup A_2$ ),  $\alpha = s_2/(s_1 + s_2)$ ,  $s_1$ ,  $s_2$  are the areas of  $A_1$ ,  $A_2$ .

Triangulating  $Q_k$  for  $n = 2$  we can thus find the  $x^{k+1}$ . For  $n > 2$  the problem of finding the center of gravity of a polyhedron becomes very cumbersome and this method is practically unfeasible in this case.

Yet the center-of-gravity method is of great theoretical interest. First, through this method it is possible to obtain a convergence rate estimate depending only on the space and the “initial uncertainty”—the quantity  $\max_{x \in Q_0} f(x) - \min_{x \in Q_0} f(x)$ —but not on individual characteristics of the function—such as its condition number. All of the estimates we have given so far do not possess these properties. In addition, for problems of small dimensionality the convergence rate is large enough. Indeed, it is seen from (11) that the accuracy of solution can be increased approximately  $e$  times in  $ne$  iterations. Thus for  $n = 10$ , to obtain a solution with accuracy to within 0.1 percent, i.e., to obtain

$$\phi_k - f^* \leq \max_{x \in Q_0} (f(x) - f^*) \cdot 10^{-3},$$

one needs to make approximately  $11 \log 10^3 \sim 190$  iterations, which is, in fact, a small number. Second, as will be shown later, this method is in some sense optimal.

## Exercises

1. Show that if  $f(x)$  is a piecewise linear function,  $I = \{k, \dots, k-m\}$  and  $m$  is sufficiently large, then method (6) is finite.
2. Show that for any function  $f(x)$  the center-of-gravity method cannot converge too fast, viz.  $v_k \geq e^{-k} v_0$ , where  $v_k$  is the volume of  $Q_k$ .

### 5.4.3 Optimal Methods

For problems of unconstrained minimization of a convex function one can define the performance of any method which uses only the subgradients and values of the function. The formulation of the next theorem is somewhat fuzzy but still obvious.

**THEOREM 3** (Nemirovskij and Yudin). For any method for minimizing the function  $f(x)$ ,  $x \in \mathbf{R}^n$ , which uses the values  $f(x)$  and  $\partial f(x)$ , we can find a convex function such that the method converges (with respect to the function) no faster than at the rate of geometric progression with ratio  $1 - c/n$ , or no faster than  $O(1/\sqrt{k})$  uniformly with respect to the dimension (here  $c$  is some absolute constant).  $\square$

We do not prove this theorem here since we would then need to give a rigorous and elaborate definition of the notion of “any method which

uses the values  $f(x)$  and  $\partial f(x)$  and make the available *a priori* information about the function more precise (initial approximation, region of localization of the minimum, the bounds for  $f(x)$  and  $\partial f(x)$ , etc.). The proof is based on the fact that for given  $x^0, \dots, x^k, f(x^0), \dots, f(x^k), \partial f(x^0), \dots, \partial f(x^k)$ , one constructs the piecewise linear function with these values of the function as well as of the subgradient at the specified points, but which maximally differs in the minimum from the quantity  $f(x)$ .

A comparison of this result with the convergence rate estimates obtained earlier leads to a major statement: there is no optimization method using the same information, i.e., the values  $f(x)$  and  $\partial f(x)$ , in which the convergence rate of the center-of-gravity method can be surpassed with respect to the order. In other words, the center-of-gravity method is optimal in some sense and any attempt to devise a more rapidly convergent method will fail.

But this method should be, however, approached with caution. To begin with, this method belongs to the wide class of “all convex functions.” In practice, one rarely has to deal with “arbitrary” convex functions. As a rule, the objective function belongs to a more narrow class, e.g., it is strongly convex, or has a sharp minimum, or has the form  $\max_{1 \leq i \leq k} f_i(x)$ , where

$f_i(x)$  are smooth, etc.). For narrow classes there are perhaps more efficient methods. Moreover, this statement is of minimax nature: there is a function that is “poor” for a given method. However, in minimizing a particular function the method may converge much faster than for the “worst” case. At the same time, the center-of-gravity method converges identically both for “good” and “poor” functions. Third, in the Nemirovskij-Yudin theorem the number of computations of the values  $f(x)$  and  $\partial f(x)$  is taken into account, ignoring the amount of computation needed to solve the concurrent auxiliary problems. It also ignores, for example, a tough job of finding the center of gravity of a polyhedron since it is not involved in the additional computations of the function and the subgradient. Indeed, one cannot regard the center-of-gravity method to be optimal; neither is it a reasonable method of optimization for  $n > 2$ . This shows that the choice of a minimization method, even given theoretical justification of its optimality (in some sense) can be complicated.

#### 5.4.4 Space Extension Methods

It is quite natural to try to modify the center-of-gravity method by eliminating its major drawbacks, viz. the laborious search of the center of gravity and the need to store the values of  $\partial f(x^i)$  obtained in the preceding iterations, retaining at the same time the rate of convergence. This can be done in the following way (Fig. 24). If a polyhedron  $Q_k$  is inside a sphere, to find the center of gravity presents no problem—it coincides with the center of the sphere. We denote this point  $x^{k+1}$  and calculate  $\partial f(x^{k+1})$ ,

we have thus “truncated” half the sphere. The other half of the sphere can be inscribed in an ellipsoid of minimal volume. By a linear space transformation we convert this ellipsoid to a sphere and reiterate the procedure. In this case there is no need to store the polyhedron  $Q_k$  proper and the constraints which define it, i.e., the  $\partial f(x^i)$ ,  $i = 0, \dots, k$ . It suffices to store at the  $k$ th step the point  $x^k$  as well as the linear space transformation defined by the matrix  $H_k$ . We have thus obtained the *ellipsoid method*:

$$\begin{aligned} x^{k+1} &= x^k - \gamma_k H_k \partial f(x^k), \\ \gamma_k &= \frac{\rho}{n+1} \left( \frac{1}{\sqrt{n^2 - 1}} \right)^k (H_k \partial f(x^k), \partial f(x^k))^{-1/2}, \\ H_{k+1} &= H_k - \frac{2}{n+1} \frac{H_k \partial f(x^k) \partial f(x^k)^T H_k}{(H_k \partial f(x^k), \partial f(x^k))}, \quad H_0 = I, \end{aligned} \quad \text{LH} \quad (12)$$

where  $\rho$  is the radius of the initial ball with center at  $x^0$  at which the minimum point is localized.

**THEOREM 4.** For method (12) the following estimate holds in the space  $\mathbf{R}^n$ ;  $n \geq 2$ :

$$\begin{aligned} \phi_k - f^* &\leq cq^k, \quad c = \max_{\|x-x^0\| \leq \rho} (f(x) - f^*), \\ q &= n(n-1)^{-(n-1)/2n} (n+1)^{-(n+1)/2n}. \end{aligned} \quad (13)$$

We omit the details of proof of this theorem. It is based on the easily verifiable fact that the volume of a minimal ellipsoid circumscribed around a hemisphere (Fig. 24) is  $2q^n$  times greater than the volume of the hemisphere. Hence, at each step the volume of the region of localization of the minimum diminishes by the factor  $q^n$ . The rest of the proof is the same as of Theorem 2.  $\square$

We can see that the behavior of method (12) is similar to that of the center-of-gravity method (convergence at the rate of geometric progression with ratio not depending on the objective function but depending on the dimension of the space). However, in the ellipsoid method the progression ratio is closer to one, i.e.,  $q \sim 1 - 1/(2n^2)$  instead of  $q \sim 1 - 1/(en)$  in method (9). For large dimensions of space, the loss in convergence rate is substantial and method (12) is no longer efficient. For example, for  $n = 10$  one needs to execute almost 200 iterations in order to increase the accuracy (for the function) by a factor of  $e$ ; for  $n = 30$  it is almost 2,000 iterations.

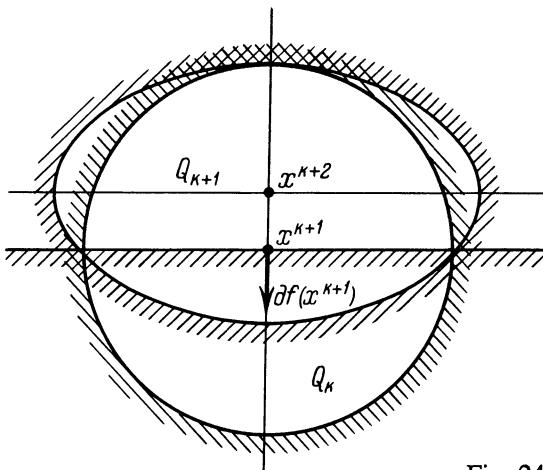


Fig. 24 The ellipsoid method.

N.Z. Shor arrived at methods similar to (12) in the different way. He suggested combining the subgradient method with the *space extension* procedure. The latter is directed either towards the last subgradient or towards the last two subgradients. The extension factor is given by a parameter which is chosen heuristically. This (see Exercises 3 and 4 below) leads to methods of the following form:

$$\begin{aligned} x^{k+1} &= x^k - \gamma_k H_k \partial f(x^k) \\ H_{k+1} &= H_k - \left(1 - \frac{1}{\alpha_k^2}\right) \frac{H_k s^k (s^k)^T H_k}{(H_k s^k, s^k)}, \quad H_0 = I, \end{aligned} \tag{14}$$

where  $\alpha_k$  is the space extension coefficient in the  $k$ th iteration,  $\gamma_k$  is the step size,  $s^k$  is the direction of extension. All these quantities can be chosen in varied ways. For example,

$$s^k = \partial f(x^k), \quad \gamma_k = \frac{2(f(x^k) - f^*)}{(H_k \partial f(x^k), \partial f(x^k))}, \quad \alpha_k = \infty, \tag{15}$$

$$s^k = \partial f(x^k), \quad \gamma_k = \lambda \frac{f(x^k) - f^*}{(H_k \partial f(x^k), \partial f(x^k))}, \quad \alpha_k = \alpha, \tag{16}$$

$$s^k = \partial f(x^k) - \partial f(x^{k-1}), \quad \gamma_k = \operatorname{argmin}_\gamma f(x^k - \gamma H_k \partial f(x^k)), \quad \alpha_k = \alpha, \tag{17}$$

where  $f^* = \min f(x)$  is assumed to be known.

It is obvious that Shor's methods are related to the variable metric methods for minimizing smooth functions, described in Section 3.3. Shor's methods can be used for nonsmooth optimization as well as the smooth optimization. The convergence of these methods is demonstrated by the following theorem.

**THEOREM 5.** Let  $f(x) = (Ax, x)/2 - (b, x)$ ,  $A > 0$ . Then methods (14), (15) and (14), (17), with  $\alpha = \infty$ , are finite:  $x^n = x^* = A^{-1}b$ .  $\square$

Little is known about the convergence and the rate of convergence of methods (14) in the general case. By the Nemirovskij-Yudin theorem, for an arbitrary convex function they cannot converge faster than at the rate of geometric progression with ratio  $1 - 1/(cn)$ . Shor analyzes a different class of functions satisfying the condition

$$N(f(x) - f^*) \leq (\partial f(x), x - x^*) \leq M(f(x) - f^*). \quad (18)$$

These functions are referred to as *approximately homogeneous* (cf. (30) of Section 3.3). It can be proved for these functions that if

$$\alpha_k \equiv \alpha = (M + N)/(M - N), \quad \lambda = 2MN/(M + N), \quad (19)$$

the method (14), (16) converges with the rate of geometric progression with ratio  $\alpha^{1/n}$ :

$$\phi_k - f^* \leq c\sqrt{k}\alpha^{-k/n}. \quad (20)$$

Thus, the closer  $M$  is to  $N$  (i.e., the closer the function is to being homogeneous), the larger  $\alpha$  and the faster the convergence. In the limit for a homogeneous function ( $M = N$ ) one can take  $\alpha = \infty$ . The method is then finite (this fact was noted for a quadratic function ( $M = N = 2$ ) in Theorem 5).

### Exercises

3. Assume that for some  $\alpha > 0$  and  $s \in \mathbf{R}^n$ ,  $\|s\| = 1$ ,  $R_\alpha(s)$  is a linear operator on  $\mathbf{R}^n$  defined by  $R_\alpha(s)x = x + (\alpha - 1)ss^T x$ . Verify that  $R_\alpha(s)$  is an operator that extends by the factor  $\alpha$  in the direction  $s$ , i.e.,  $R_\alpha(s)s = \alpha s$ ,  $R_\alpha(s)x = x$  for  $(x, s) = 0$ .
4. Show that  $H_k$  in (14) is the result of successive applications of extension operators, i.e.,  $H_k = P_k P_k^T$ ,  $P_0 = I$ ,  $P_{i+1} = P_i R_{\alpha_i^{-1}}(s^i)$ .
5. Verify that for  $\alpha = 1$  (i.e., without extension) method (14), (16) with  $\lambda = 1$  becomes the subgradient method (7) of Section 5.3.
6. Prove Theorem 6 and compare it with the results of Section 3.3. What does the choice  $\alpha_k = \infty$  imply?

## 5.5 THE INFLUENCE OF NOISE

### 5.5.1 The Statement of the Problem

Let us examine now the behavior of the subgradient method for minimizing a convex function  $f(x)$  on  $\mathbf{R}^n$  in noise.

Let

$$x^{k+1} = x^k - \gamma_k s^k, \quad s^k = \partial f(x^k) + r^k, \quad (1)$$

where  $r^k$  is the noise imposed on the subgradient. This noise can be of different kinds, e.g., inaccuracy in computation, errors in measurements, approximate formulas, and the like. Formally speaking, noise can be absolute or relative, or deterministic or random. We will examine most typical kinds of noise. We are interested in the convergence, estimation of the rate of convergence, as well as rational techniques for choosing the  $\gamma_k$ , i.e., the same problems as those solved in Chapter 4 for the smooth case.

### 5.5.2 Absolute Deterministic Noise

Suppose the errors in computing the subgradient satisfy the condition

$$\|r^k\| \leq \varepsilon, \quad (2)$$

where  $r^k$  is the absolute level of noise. As was shown, in smooth problems this kind of noise violates the convergence—the gradient method converges only into a neighborhood of the minimum, the size of which depends on  $\varepsilon$  as well as on the condition number of the problem. For nonsmooth problems the situation is different: for a low noise level, in the case of a sharp minimum the convergence will remain unchanged if the  $\gamma_k$  is chosen in a particular way. This is due to the fact that  $\partial f(x)$  does not tend to 0 while approaching the sharp minimum.

**THEOREM 1.** Let  $f(x)$  be a convex function on  $\mathbf{R}^n$  and let  $x^*$  be a sharp minimum point of  $f(x)$ , i.e.,  $f(x) - f(x^*) \geq \alpha \|x - x^*\|$ ,  $\alpha > 0$ . Let  $\varepsilon < \alpha$  in (2). Then for any  $x^0$  there are  $\gamma_0 > 0$ ,  $q < 1$ , such that in method (1)  $\gamma_k = \gamma_0 q^k$  one has

$$\|x^k - x^*\| \leq \|x^0 - x^*\| q^k. \quad \square \quad (3)$$

To make use of the method in Theorem 1 for choosing the step size, one needs to have the detailed information about the problem (to have estimates for  $L$ ,  $\alpha$ ,  $\varepsilon$ ,  $\|x^0 - x^*\|$ ). Without this information, the incorrect choice of the  $\gamma_0$  and  $q$  may result in the situation that the method stops outside the minimum point. We will not go into discussion of other, more

realistic methods for adjusting the step size; of greater importance is the fact that the convergence of the subgradient method at the rate of geometric progression is theoretically feasible for nonsmooth problems in absolute noise.

### 5.5.3 Relative Deterministic Noise

Suppose the relative noise level is given:

$$\|r^k\| \leq \alpha \|\partial f(x^k)\|. \quad (4)$$

In smooth problems the method converges for any  $\alpha < 1$  (Theorem 2 in Section 4.2). Nonsmooth functions, again, make the situation different. Let us briefly analyze the convergence. The pseudogradient condition of algorithm (1) relative to the Lyapunov function

$$V(x) = \|x - x^*\|^2/2 \quad (5)$$

has the form  $(s^k, x^k - x^*) \geq 0$ . But

$$(s^k, x^k - x^*) = (\partial f(x^k) + r^k, x^k - x^*) \geq (\cos \phi_k - \alpha) \|\partial f(x^k)\| \|x^k - x^*\|,$$

where  $\phi_k$  is the angle between the  $\partial f(x^k)$  and  $x^k - x^*$ ,  $0 \leq \phi_k \leq \pi/2$ . Hence, if

$$0 \leq \phi_k \leq \phi < \pi/2, \quad \alpha \leq \cos \phi, \quad (6)$$

then the pseudogradient conditions is satisfied. Condition (6) is substantially more constraining than the condition  $\alpha < 1$ . For worse ill-conditioned functions the  $\cos \phi$  is smaller and the method is more sensitive to relative noise. Figure 18 shows that even a small error in determining the direction of the subgradient can make the method fail in approaching the minimum point. For this very reason, a generalization (which seems to be natural) of the subgradient method

$$x^{k+1} = x^k - \gamma_k H \partial f(x^k), \quad (7)$$

$H > 0$  being some matrix, may not converge at all.

### 5.5.4 Absolute Random Noise

Suppose that different kinds of the noise  $r^k$  are random, mutually independent, centered, and have bounded variance:

$$E r^k = 0, \quad E \|r^k\|^2 \leq \sigma^2. \quad (8)$$

**THEOREM 2.** Let  $f(x)$  be a convex function and let  $\|\partial f(x)\| \leq c$  for all  $x$ . Also, let there be a minimum point  $x^*$ , let (8) hold, and let

$$\sum_{k=0}^{\infty} \gamma_k = \infty, \quad \sum_{k=0}^{\infty} \gamma_k^2 < \infty. \quad (9)$$

Then in method (1)

$$\min_{0 \leq i \leq k} f(x^i) \rightarrow f(x^*) \text{ a.s. } \square$$

Thus, as in the smooth case, the method converges in additive random noise of any level if the  $\gamma_k$  satisfies (9). The difference between the smooth case and the nonsmooth case lies in the fact that in the smooth case the noise makes it necessary to change the method for adjusting the step size (one needs to choose  $\gamma_k \rightarrow 0$  instead of  $\gamma_k \equiv \gamma$ ), whereas in the nonsmooth case the noise has little effect (noise, or no noise, one needs to take  $\gamma_k \rightarrow 0$ ). It is not quite clear what the situation is with the convergence under condition (8). If  $f(x)$  is strongly convex, then taking  $\gamma_k = \gamma/k$  for sufficiently large  $\gamma$  one can obtain a convergence of the order  $O(1/k)$ ; the proof is routine. However, for a sharp minimum, what is more typical for nonsmooth problems, the question of the convergence rate has not been studied enough.

### Exercise

1. Show that if  $f(x)$  has a sharp minimum with constant  $\ell$ , then for all  $x$  in the region  $S = \{x: \|x - x^*\| \leq \rho\}$  we have the inequality

$$(\partial f(x), x - x^*) \geq (\ell/L) \|\partial f(x)\| \|x - x^*\|,$$

where  $L = \max_{\underline{x} \in S} \|\partial f(x)\|$ , i.e., (6) holds for  $\cos \phi = \ell/L$ .

## 5.6 SEARCH METHODS

Let us examine the problem of minimizing a convex function  $f(x)$  in the situation where the values of  $f(x)$  at an arbitrary point is the only available information about the function.

### 5.6.1 The One-dimensional Case

Search for the minimum of a one-dimensional convex function  $f(x)$  on the segment  $[a, b] \subset \mathbf{R}^1$  is easy if one follows the following geometrically obvious arguments (Fig. 25(a)). If the values of  $f(x)$  are computed at

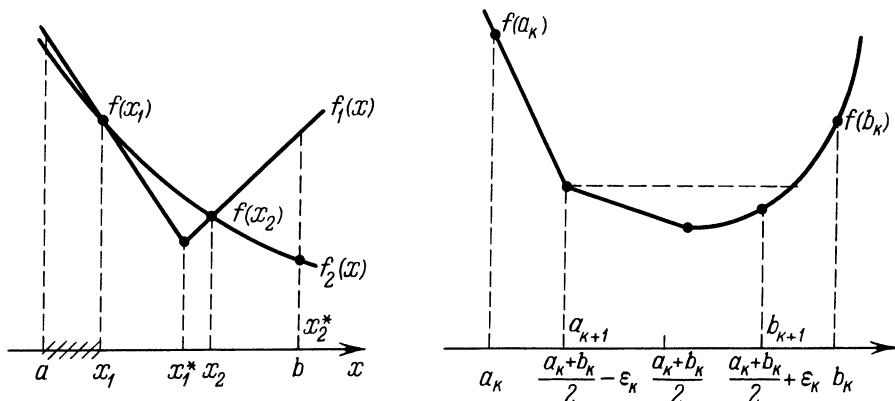
two points  $x_1, x_2$ ,  $a < x_1 < x_2 < b$ , then a minimum point  $x^*$  cannot lie on the segment  $[a, x_1]$  if  $f(x_1) > f(x_2)$ , nor on the segment  $[x_2, b]$  if  $f(x_2) > f(x_1)$  (if  $f(x_1) = f(x_2)$ , then one of the minimum points belongs to  $[x_1, x_2]$ ). Hence, upon computation of the two values of the function the region of localization of the minimum can be reduced. The simplest algorithm which implements this idea arranges points on each segment symmetrically with respect to its center (Fig. 25(b)):

$$\begin{aligned}
 a_0 &= a, \quad b_0 = b, \quad \varepsilon_k = \alpha(b_k - a_k)/2, \quad 0 < \alpha < 1, \\
 a_{k+1} &= \begin{cases} a_k & \text{if } f((a_k + b_k)/2 - \varepsilon_k) < f((a_k + b_k)/2 + \varepsilon_k), \\ (a_k + b_k)/2 - \varepsilon_k & \text{otherwise,} \end{cases} \\
 b_{k+1} &= \begin{cases} b_k & \text{if } f((a_k + b_k)/2 - \varepsilon_k) > f((a_k + b_k)/2 + \varepsilon_k), \\ (a_k + b_k)/2 + \varepsilon_k & \text{otherwise,} \end{cases} \\
 a_{k+1} &= (a_k + b_k)/2 - \varepsilon_k, \\
 b_{k+1} &= (a_k + b_k)/2 + \varepsilon_k, \\
 &\text{if } f((a_k + b_k)/2 - \varepsilon_k) = f((a_k + b_k)/2 + \varepsilon_k).
 \end{aligned} \tag{1}$$

Obviously,

$$0 \leq b_{k+1} - a_{k+1} \leq (1 + \alpha)(b_k - a_k)/2,$$

Fig. 25 One-dimensional search.



so that the length of the segment on which a minimum is localized is reduced in each iteration roughly to half if  $\alpha$  is small. Clearly, for  $\alpha \ll 1$ , (1) is merely a difference analog of the dichotomy method (Section 5.3).

Of greater advantage yet is to use the preceding values of the function (one of those on the segment  $[a_{k+1}, b_{k+1}]$  was found in the preceding iteration). In that case if  $\alpha$  is chosen from the relation

$$(1 + \alpha)/2 = \beta, \quad \beta^2 = 1 - \beta, \quad \beta = (\sqrt{5} - 1)/2 \quad (2)$$

(the equation of the “golden section” of the segment), then one of the points  $(a_{k+1} + b_{k+1})/2 \pm \varepsilon_{k+1}$  will coincide with the  $(a_k + b_k)/2 \mp \varepsilon_k$ , i.e., each iteration requires only a single computation of the function. In the bisection method ((1) with  $\alpha \ll 1$ ) the segment reduces by the factor  $\sqrt{2} \approx 1.41$  per a single computation, whereas in the golden-section method (1), (2) it reduces by the factor  $2/(1+\alpha) = (\sqrt{5} + 1)/2 \approx 1.62$ , which is somewhat better.

Yet, even of greater advantage is to make  $\alpha$  be dependent on  $k$ . This is exactly what has been done in *Fibonacci's method*, described in detail in, for example, [0.2, 0.8, 0.18]. It is not hard to see that all of the foregoing methods search for the minimum of a convex function as well as any *unimodal* function (i.e., such that the local minimum coincides with the global minimum). Fibonacci's method can be shown to reduce the length of the localization segment per a single computation of the function maximally fast, viz. it is optimal in the minimax sense in the class of unimodal functions. Nevertheless Fibonacci's method is used rarely because: (1) it is only insignificantly superior to the golden-section method, at the same time it involves additional computation in order to construct new points; (2) it requires that the number of iterations be determined in advance. Since the natural criterion for a termination of a one-dimensional minimization process is not the dimension of the region of localization of the minimum but, instead, the proximity of the resulting value to the minimal value of the function, it is not easy to determine in advance the number of steps needed; and (3) it is optimal only in the minimax sense, i.e., with a view of the “worst” unimodal function. A faster convergence for concrete functions may be provided by other methods.

This should suffice to demonstrate how cautious one has to be in treating the theoretical conclusions concerning the optimality of the methods (see Section 4.3).

## 5.6.2 The Multidimensional Case

Most ideas underlying search methods for minimizing smooth functions (Section 3.4) do not carry over to the nonsmooth case. Thus methods of successive one-dimensional minimization such as coordinatewise descent, as

we have already seen (Fig. 18) may not converge for nondifferentiable functions. The ideas of local linear or quadratic approximation of the objective function are also not effective. On the other hand, the subgradient method (Section 5.2) and generalizations of it (5.3) cannot be applied if the subgradient is replaced by its finite-difference approximation—we have already observed (Section 5.4) that the subgradient method is generally unstable under deterministic errors. Finally, the one-dimensional search method described above does not carry over simply to the multidimensional case. The point is that having calculated the function at several points, it is difficult to localize the region of the minimum in the multidimensional case. Due to the above-indicated difficulties there are comparatively few theoretically investigated and justified search methods for minimizing nonsmooth functions.

Let us describe one of them, the idea of which is very simple and instructive. For the problem of minimizing a convex function  $f(x)$  on  $\mathbf{R}^n$  it has the form

$$\begin{aligned} x^{k+1} &= x^k - \gamma_k s^k, \\ s^k &= \delta_k^{-1} [f(x^k + \alpha_k g^k + \delta_k h^k) - f(x^k + \alpha_k g^k)] h^k, \end{aligned} \quad (3)$$

where  $g^k, h^k$  are independent random vectors uniformly distributed on the cube  $Q = \{x: |x_i| \leq 1, i = 1, \dots, n\}$ ,  $\alpha_k, \delta_k, \gamma_k$  are certain scalar sequences. In other words, the step of random search is made (in the direction  $h^k$ ), not from the point  $x^k$ , but rather from the “randomized” point  $x^k + \alpha_k g^k$ . Owing to the introduction of such a *randomization*, there occurs a smoothing of the initial function. One can show that

$$E(s^k | x^k) = c \nabla f(x^k, \alpha_k) + \beta_k, \quad \|\beta_k\| \leq c_1 \delta_k / \alpha_k, \quad (4)$$

where  $f(x, \alpha)$  is the *smoothed* function,

$$f(x, \alpha) = \frac{1}{(2\alpha)^n} \int_Q f(x + \alpha y) dy, \quad (5)$$

and  $f(x, \alpha)$  is a convex differentiable function whose gradient satisfies a Lipschitz condition with constant  $c\sqrt{n}/\alpha_k$ . Thus (3) can be viewed as a gradient method of minimizing the smoothed function (5) in the presence of noise. By regulating the smoothing coefficient  $\alpha_k$ , the size of the trial step  $\delta_k$  and of the working step  $\gamma_k$ , one can get the method to converge. Thus, if

$$\sum_{k=0}^{\infty} \gamma_k = \infty, \quad \sum_{k=0}^{\infty} \gamma_k^2 < \infty, \quad \gamma_k / \alpha_k \rightarrow 0, \quad (6)$$

$$\delta_k / \alpha_k \rightarrow 0, \quad \alpha_k \rightarrow 0, \quad |\alpha_k - \alpha_{k+1}| / \gamma_k \rightarrow 0,$$

then the method converges with probability 1 to the set of minimum points (if the latter is nonempty). Similarly, the procedure of smoothing by means of randomization can be applied for constructing other methods.

Of course, the convergence rate of method (3) is very low. The problem of constructing effective search methods for minimizing nonsmooth convex functions in the multidimensional case remains an open question.

## CHAPTER 6

### SINGULARITY, MULTIMODALITY, NONSTATIONARITY

In practice the engineer rarely meets with the ideal situation similar to that described in Chapters 1 and 3. We have discussed a few of the complications—such as noise and nondifferentiability. We consider now other kinds of factors complicating the solution of problems of unconstrained minimization, viz. singularity of the minimum, multimodality and nonstationarity. We examine the behavior of standard methods in such situations, and investigate specific techniques to overcome the difficulties.

#### 6.1 A SINGULAR MINIMUM

In Chapters 1 and 3 we studied optimization methods primarily for the case of a nonsingular minimum (i.e., under the assumption that at the minimum point  $x^*$ ,  $\nabla^2 f(x^*) > 0$ ). In the ensuing discussion we drop this assumption.

##### 6.1.1 The Behavior of Standard Methods

Let us examine the behavior of the simplest gradient method of unconstrained minimization of a differentiable function  $f(x)$ :

$$x^{k+1} = x^k - \gamma \nabla f(x^k) \quad (1)$$

in the situation where the nonsingularity of the minimum point is not assumed, but  $f(x)$  is convex. We have seen (Theorem 1 of Section 1.4) that

under minimal assumptions we have  $\nabla f(x^k) \rightarrow 0$  for (1). Thus, for convex functions a stronger result holds true.

**THEOREM 1.** Let  $f(x)$  be a convex differentiable function in  $\mathbb{R}^n$  whose gradient satisfies a Lipschitz condition with constant  $L$ , and the set of minimum points  $X^* = \underset{x \in \mathbb{R}^n}{\operatorname{Argmin}} f(x)$  is nonempty. Then method (1) with  $0 < \gamma < 2/L$  converges to some point  $\tilde{x} \in X^*$ ,  $f(\tilde{x}) = f^*$ , with

$$f(x^k) - f^* = o(1/k). \quad (2)$$

**PROOF.** We use the inequality (Lemma 2 of Section 1.4) which holds for convex functions whose gradient satisfies a Lipschitz condition with constant  $L$ . Then  $(\nabla f(x), x - \hat{x}) \geq L^{-1} \|\nabla f(x)\|^2$ , where  $\hat{x}$  is an arbitrary minimum point. Hence

$$\begin{aligned} \|x^{k+1} - \hat{x}\|^2 &= \|x^k - \hat{x}\|^2 - 2\gamma(\nabla f(x^k), x^k - \hat{x}) + \gamma^2 \|\nabla f(x^k)\|^2 \\ &\leq \|x^k - \hat{x}\|^2 - \gamma(2/L - \gamma) \underbrace{\|\nabla f(x^k)\|^2}_{\|x^k - \hat{x}\|^2}. \end{aligned} \quad (3)$$

Summing over  $k$ , we obtain that for  $0 < \gamma < 2/L$

$$\sum_{k=0}^{\infty} \|\nabla f(x^k)\|^2 < \infty, \quad (4)$$

i.e.,  $\nabla f(x^k) \rightarrow 0$ . The sequence  $x^k$  is bounded since  $\|x^k - \hat{x}\| \leq \|x^0 - \hat{x}\|$ . It is therefore possible to choose the convergent subsequence  $x^{k_i} \rightarrow \tilde{x}$ . By the continuity of  $\nabla f(x)$  we have here  $\nabla f(\tilde{x}) = 0$ , i.e.,  $\tilde{x} \in X^*$ . Replacing the  $\hat{x}$  with  $\tilde{x}$  in (3) yields  $x^k \rightarrow \tilde{x}$ .

Next we estimate the rate of convergence with respect to the function. We have (see (9) in Section 1.4)

$$f(x^{k+1}) \leq f(x^k) - \alpha \|\nabla f(x^k)\|^2, \quad \alpha = \gamma(1 - L\gamma/2) > 0.$$

From the convexity of  $f(x)$  we have

$$f(x): f(x^k) - f^* \leq (\nabla f(x^k), x^k - \tilde{x}) \leq \|\nabla f(x^k)\| \|x^k - \tilde{x}\|.$$

Hence for  $u_k = f(x^k) - f^*$  we obtain  $u_{k+1} \leq u_k - \alpha \|x^k - \tilde{x}\|^{-2} u_k^2$ , and applying Lemma 6 of Section 2.2 we have

$$u_k \leq \left( \frac{1}{u_0} + \alpha \sum_{i=0}^{k-1} \|x^i - \tilde{x}\|^{-2} \right)^{-1},$$

$$ku_{k+1} \leq \left( \frac{1}{u_0 k} + \frac{\alpha}{k} \sum_{i=0}^{k-1} \|x^i - \tilde{x}\|^{-2} \right)^{-1}.$$

Since according to what has been proved  $\|x^i - \tilde{x}\| \rightarrow 0$  as  $i \rightarrow \infty$ , we have:

$$\|x^i - \tilde{x}\|^{-2} \rightarrow \infty \quad \text{and} \quad k^{-1} \sum_{i=0}^{k-1} \|x^i - \tilde{x}\|^{-2} \rightarrow \infty \quad \text{as } k \rightarrow \infty.$$

Therefore the right-hand side of the last inequality tends to zero as  $k \rightarrow \infty$ . This implies  $u_k = o(1/k)$ .  $\square$

We note that it is impossible to single out in advance the point  $\tilde{x} \in X^*$  to which  $x^k$  converges. For example, this point can vary for varied  $\gamma$  (for fixed  $x^0$ ) and does not necessarily coincide with  $x^*$ , which is the point of  $X^*$  closest to  $x^0$ . However,  $\tilde{x}$  cannot lie too far away from  $x^*$ . Indeed, substituting  $x^*$  for  $\tilde{x}$  in (3), we obtain  $\|x^k - x^*\| \leq \|x^0 - x^*\|$ , i.e.,

$$\|\tilde{x} - x^*\| \leq \|x^0 - x^*\| = \rho(x^0, X^*). \quad (5)$$

It follows from Theorem 1 that the gradient method converges (in the convex case) without any assumptions on the nonsingularity of the minimum. In this case, the convergence rate of order  $o(1/k)$  with respect to the function is guaranteed. However, the convergence rate with respect to the variables can be substantially lower. For example, let  $f(x) = p^{-1}|x|^p$  for  $|x| \leq 1$ ,  $f(x) = |x|$  for  $|x| > 1$ ,  $p > 2$ ,  $x \in \mathbf{R}^1$ . Then  $f(x)$  satisfies the conditions of Theorem 1,  $x^* = 0$ , and from (1) we have

$$|x^{k+1}| = |x^k - \gamma(x^k)^{p-1}| \quad \text{for } |x^0| \leq 1.$$

Using the result of Exercise 3 of Section 2.2, we find that for  $0 < \gamma < 2$  one has  $|x^k| = O(k^{-1/(p-2)})$ . Thus, for a sufficiently large  $p$ , for any  $\alpha > 0$  we can find a function  $f(x)$  such that the gradient method converges more slowly than  $k^{-\alpha}$ . Note that for the same case one has  $f(x^k) = O(k^{-p/(p-2)})$ , which corresponds to estimate (2) and shows that it is impossible to improve it.

Let us show now that no convergence rate with respect to the variable can be guaranteed under the conditions of Theorem 1. Indeed, for any  $\gamma > 0$ ,  $\varepsilon_k > 0$ ,  $\varepsilon_k \rightarrow 0$  we construct the convex function  $f(x)$ ,  $x \in \mathbf{R}^n$  with the single minimum point  $x^* = 0$ , with the derivative  $f'(x)$  satisfying a Lipschitz condition with constant  $L = 1/\gamma$  such that the estimate

$$|x^k| \geq \varepsilon_k \quad \forall k$$

holds for method (1) applied to  $f(x)$ . Suppose that  $\varepsilon_k$  and  $\delta_k = \varepsilon_k - \varepsilon_{k+1}$  are monotone decreasing (otherwise we construct  $\tilde{\varepsilon}_k \geq \varepsilon_k$ ,  $\tilde{\varepsilon}_k$  having the desired property). Define a function  $g(x)$ :  $g(\varepsilon_k) = \delta_k$ ;  $g(x)$  is linear on  $[\varepsilon_{k+1}, \varepsilon_k]$ ,  $g(0) = 0$ ,  $g(x) = \delta_0$  for  $x \geq \varepsilon_0$ ,  $g(x) = -g(-x)$  for  $x < 0$ . The function  $g(x)$  is defined on  $\mathbf{R}^1$ ; it is monotone nonincreasing and satisfies a Lipschitz condition with constant 1. The function

$$f(x) = (1/\gamma) \int_0^x g(t) dt$$

is the required one: it is differentiable,  $f'(x) = (1/\gamma)g(x)$ , and convex,  $f(0) = 0$ ,  $f(x) > 0$  for  $x \neq 0$ ,  $f'(x)$  satisfies a Lipschitz condition with constant  $1/\gamma$ . If we take  $x^0 \geq \varepsilon_0$ , it is not hard to prove by induction that  $x^k \geq \varepsilon_k$  for all  $k$ , where  $x^k$  are the points generated by method (1).

Thus, the gradient method may converge as slowly as possible with respect to the variable in the case of nonquadratic functions with single minimum points.

Let us now analyze more closely the behavior of the gradient method for a quadratic function:

$$f(x) = (Ax, x)/2 - (b, x), \quad A > 0. \quad (6)$$

Although the problem of minimizing  $f(x)$  is a nonsingular one (since  $A > 0$ , then the minimum point  $x^*$  exists, is unique, globally stable, and  $f(x)$  is strongly convex), we are interested in the case of an ill-conditioned problem, which in some sense is close to a singular one. Let  $L$  and  $\ell$  be the largest and the smallest eigenvalues of  $A$ ,  $\mu = L/\ell \gg 1$ . As we know (Theorem 3 of Section 1.4), for the choice  $\gamma = 2/(L + \ell)$  (this is the best choice) for the gradient method (1) we have the estimate

$$\|x^k - x^*\| \leq \|x^0 - x^*\| q^k, \quad q = (L - \ell)/(L + \ell) = (\mu - 1)/(\mu + 1),$$

this estimate being unimprovable (see the examples after that theorem). Since

$$\begin{aligned} 2(f(x^k) - f^*) &= (A(x^k - x^*), x^k - x^*) \leq \|A\| \|x^k - x^*\|^2 \\ &\leq L \|x^0 - x^*\|^2 q^{2k}, \end{aligned}$$

it is possible to guarantee the convergence with respect to the function at the rate of geometric progression with ratio  $q_1 = q^2$ . However, for ill-conditioned problems  $\mu \gg 1$ , and  $q_1 \approx 1 - 4/\mu$  is very close to 1. It is therefore possible to obtain an estimate of the convergence rate with respect to the function which does not depend on the condition number of the problem.

**THEOREM 2.** Method (1) for minimizing (6) for  $0 < \gamma < 2/L$  converges to  $x^*$ , and for sufficiently large  $k$  we have

$$f(x^k) - f^* \leq \frac{\|x^0 - x^*\|^2}{2\gamma(2k+1)} \left(1 - \frac{1}{2k+1}\right)^{2k} < \frac{\|x^0 - x^*\|^2}{4\gamma ek}. \quad (7)$$

**PROOF.**

$$\begin{aligned} x^k - x^* &= (I - \gamma A)^k (x^0 - x^*) , \\ 2(f(x^k) - f^*) &= (A(I - \gamma A))^{2k} (x^0 - x^*, x_0 - x^*) \\ &\leq \|x^0 - x^*\|^2 \|A(I - \gamma A)^{2k}\| \\ &\leq \|x^0 - x^*\|^2 \max_{0 \leq \lambda \leq L} |\lambda(1 - \gamma\lambda)^{2k}| \\ &\leq \|x^0 - x^*\|^2 \max_{0 \leq \lambda \leq L} \phi(\lambda) , \end{aligned}$$

where  $\phi(\lambda) = \lambda(1 - \gamma\lambda)^{2k}$ . Since the roots of  $\phi'(\lambda)$  are  $\lambda_1 = 1/\gamma$  and  $\lambda_2 = 1/(\gamma(2k+1))$  and  $\phi(\lambda_1) = 0$ ,  $\phi(0) = 0$ , then the maximum of the  $\phi(\lambda)$  on  $[0, \underline{L}]$  can be obtained either for  $\lambda = \lambda_2$ , or for  $\lambda = L$ . Since

$$\begin{aligned} \phi(\lambda_2) &= \frac{1}{\gamma(2k+1)} \left(1 - \frac{1}{2k+1}\right)^{2k} < \frac{1}{2\gamma ek} . \\ \phi(L) &= L(1 - \gamma L)^{2k} , \quad \text{while } |1 - \gamma L| < 1 , \end{aligned}$$

then for sufficiently large  $k$  we have  $\max_{0 \leq \lambda \leq L} \phi(\lambda) = \phi(\lambda_2)$ , yielding (7).  $\square$

Therefore, it is possible to guarantee an estimate similar to  $f(x^k) - f^* \leq c/k$ , where the constant  $c$  does not depend on the condition number.

For the convergence rate with respect to the argument, no estimate which is “uniform with respect to the condition number” is possible. Indeed, for any  $0 < \alpha < 1$  and any  $k$  one can construct a quadratic function of the form (6) as well as an initial approximation  $x^0$  such that  $\|x^k - x^*\| > \alpha \|x^0 - x^*\|$  for method (1) for any  $\gamma$ . Moreover, it suffices in this case to take  $n = 2$ , the set of such points  $x^0$  is sufficiently “extended.”

We proceed now to analyze a standard method of minimization, viz. the conjugate gradient method (Section 3.2). So far the behavior of this method for a singular minimum has not been studied for the general case; apparently, the major advantage of this method—its rapid convergence—is gone. We will consider only the case of a quadratic function (6), assuming that the problem is of large dimension (hence we are not able to take advantage

of the finiteness property of the method). In (30) of Section 3.2 we found an estimate of the convergence rate of the method:

$$\|x^k - x^*\| \leq 2(L/\ell)^{1/2} \|x^0 - x^*\| q^k, \quad q = (\sqrt{\mu} - 1)(\sqrt{\mu} + 1),$$

where the progression ratio  $q$  depends on the condition number and is near 1 for ill-posed problems. As earlier, we can obtain a convergence rate estimate with respect to the function not depending on the condition number.

**THEOREM 3.** In the conjugate gradient method, for the function (6) we have the estimate

$$f(x^k) - f^* \leq \frac{L \|x^0 - x^*\|^2}{2(2k+1)^2}. \quad (8)$$

**PROOF.** By (27) of Section 3.2, we have

$$x^k - x^* = P_k(A)(x^0 - x^*),$$

where  $P_k(\lambda)$  is a polynomial of degree  $k$  possessing the property

$$\begin{aligned} 2(f(x^k) - f^*) &= (AP_k(A)^2(x^0 - x^*), x^0 - x^*) \\ &= \min_{R \in \mathcal{R}} (AR(A)^2(x^0 - x^*), x^0 - x^*), \end{aligned}$$

where  $\mathcal{R}$  is the set of polynomials  $R(\lambda)$  of degree  $k$  satisfying the condition  $R(0) = 1$ . Set

$$R^*(\lambda) = \frac{T_{2k+1}(\sqrt{\lambda}/\sqrt{L})}{(2k+1)'(\sqrt{\lambda}/\sqrt{L})},$$

where  $T_k(x) = \cos(k \arccos x)$  is the Chebyshev polynomial. Since  $T_{2k+1}(x)$  contains only odd powers of  $x$ , then  $R_0(x) = T_{2k+1}(\sqrt{x})/\sqrt{x}$  is a polynomial of degree  $k$  in  $x$ ,  $R_0(0) = T'_{2k+1}(0) = 2k+1$ . Hence  $R^*(\lambda) \in \mathcal{R}$ . Thus

$$\begin{aligned} 2(f(x^k) - f^*) &\leq (AR^*(A)^2(x^0 - x^*), x^0 - x^*) \\ &\leq \|x^0 - x^*\|^2 \max_{0 \leq \lambda \leq L} |\lambda R^*(\lambda)^2| \\ &= \frac{L \|x^0 - x^*\|^2}{(2k+1)^2} \max_{0 \leq \lambda \leq L} \left| T_{2k+1}\left(\frac{\sqrt{\lambda}}{\sqrt{L}}\right) \right| = \frac{L \|x^0 - x^*\|^2}{(2k+1)^2}, \end{aligned}$$

since  $\max_{0 \leq x \leq 1} |T_k(x)| = 1$ .  $\square$

We see that regardless of the condition number of the problem the conjugate-gradient method guarantees a sufficiently high rate of convergence with respect to a function of the type  $O(k^{-2})$  instead of the type  $O(k^{-1})$ , as the case is in the gradient method. Estimate (8) cannot be strengthened. Thus for any  $k$  it is possible to construct a quadratic function in the space  $\mathbf{R}^n$ ,  $n = k + 1$ , and also find an  $x^0$  such that

$$f(x^k) - f^* = L \|x^0 - x^*\|^2 / (2(2k + 1)^2).$$

Furthermore, it can be shown that any method for minimizing quadratic functions which uses the information restricted to the gradients fails to yield a higher convergence rate than that in (8), uniformly with respect to the dimension as well as over the entire class of quadratic functions  $f(x)$ .

For the convergence rate of this method with respect to the argument no estimates can be obtained which do not depend on the condition number of the problem.

The conjugate gradient method has a convergence rate of the type  $O(k^{-2})$  with respect to the function in the quadratic case (see (8)). It is useful to construct a minimization method for nonquadratic functions having the same property. Such method was suggested recently by Yu.E. Nesterov:

$$\begin{aligned} x^k &= y^k - \gamma \nabla f(y^k), & \gamma = 1/L, \quad y^1 = x^0, \\ y^{k+1} &= x^k + \beta_k(x^k - x^{k-1}), & \beta_k = (\alpha_k - 1)/\alpha_{k+1}, \\ \alpha_{k+1} &= (\sqrt{4\alpha_k^2 + 1} + 1)/2, & \alpha_1 = \frac{1}{2}. \end{aligned} \tag{9}$$

Thus, the method generates two sequences  $x^k$  and  $y^k$ ; each iteration requires a single computation of the gradient. If  $f(x)$  is convex,  $\nabla f(x)$  satisfies a Lipschitz condition with constant  $L$ ,  $X^*$  is nonempty and  $x^* \in X^*$ , then for method (9) we have

$$f(x^k) - f^* \leq 2Lk^{-2} \|x_0 - x^*\|^2, \tag{10}$$

i.e., the bound  $O(k^{-2})$  holds true.

It is interesting to compare these results with the estimates of the convergence rate with respect to the function obtained earlier for the nonsmooth case. For the subgradient method in the form (7) of Section 5.3, we have the result on convergence (Theorem 2 of Section 5.3) similar to Theorem 1. The rate of convergence is, however, lower than that in the smooth case: by (8) of Section 5.3,  $f(x^k) - f^*$  decreases like  $o(k^{-1/2})$ . For the ellipsoid method (12) of Section 5.4, we proved the linear convergence with respect to the function (Theorem 4 of Section 5.4), the progression ratio depending on the dimension and not depending on the condition

number of the problem or any other factors. Since for the smooth singular case we observed no linear convergence rate, one might expect that in problems of small dimension the ellipsoid method is adequate as well for minimizing smooth ill-conditioned functions.

To conclude, we comment on the behavior of Newton's method in the singular case. First of all, this method is not always correctly defined since the matrix  $\nabla^2 f(x^k)$  may turn out to be singular in an arbitrarily small neighborhood of  $x^*$ . Hence the method cannot be used to solve singular problems. There is a narrow class of problems free from this drawback. Let, say,  $\nabla^2 f(x) > 0$  for all  $x \neq x^*$  in a neighborhood of  $x^*$ , and let the matrix  $\nabla^2 f(x^*) \geq 0$  have no inverse at the point  $x^*$ . Under some additional assumptions the Newton method converges, the convergence rate being, however, substantially lower than that for the nonsingular case. For example, let  $f(x) = |x|^p$ ,  $p > 2$ ,  $x \in \mathbf{R}^1$ . Then  $f'(x) = p|x|^{p-1} \operatorname{sign} x$ ,  $f''(x) = p(p-1)|x|^{p-2}$ ,  $f''(x) > 0$  for  $x \neq 0$  and  $f''(x^*) = f''(0) = 0$ . Newton's method for  $x_0 > 0$  takes the form  $x_{k+1} = x_k - (p-1)^{-1}x_k = qx_k$ ,  $q = (p-2)/(p-1) < 1$ . Hence  $x_k = q^k x_0$ , i.e., Newton's method converges with the rate of geometric progression with ratio close to 1 for large  $p$ . Of course this is far worse than the quadratic convergence for the nonsingular case. In other situations (see Exercise 4 below) the convergence rate may be even smaller.

To summarize, the standard minimization methods remain convergent when the singular minimum of a smooth convex function is sought. In this situation, however, the rate of convergence becomes worse, sometimes drastically.

### Exercises

- For a fixed number of steps  $k$ , find the best way of choosing the parameter  $\gamma$  in method (1) for minimizing (6), starting with the estimates derived in proving Theorem 2.

*Hint:* Take  $\gamma$  such as to minimize the  $\max_{0 \leq \lambda \leq L} \lambda(1-\lambda)^{2k}$  for known  $k, L$ .

- Consider the gradient method of the form  $x^{k+1} = x^k - \gamma_k \nabla f(x^k)$  for minimizing (6) and choose (for a fixed number of steps  $k$  and the known constant  $L$ )  $\gamma_i$ ,  $0 \leq i \leq k-1$ , so that the same estimates for  $f(x^k) - f^*$  hold as in the conjugate-gradient method.

*Hint:* Solve the problem of minimizing the quantity

$$\max_{0 \leq \lambda \leq L} \left| \lambda \prod_{i=0}^{k-1} (1 - \gamma_i \lambda)^2 \right|$$

with respect to  $\gamma_i$ ,  $0 \leq i \leq k$ ,

3. Consider the case of quadratic  $f(x)$  with  $A \geq 0$  and nonempty set of minimum points. Show that all of the results concerning the convergence and the rate of convergence derived in this section for  $A > 0$  will hold.

4. Analyze the convergence rate of Newton's method for the function  $f(x) = \exp(-x^2)$ ,  $x \in \mathbb{R}^1$ , in a neighborhood of the minimum point  $x^* = 0$ . TU

### 6.1.2 Special Methods for Singular Problems

1. The regularization method. Suppose the problem of minimizing a convex function  $f(x)$  is ill-conditioned, for example, it has a singular minimum. This problem can be slightly modified if we add to  $f(x)$  a "good" function  $g(x)$  with a small "weight." In finding the minimum point of the "improved" function  $f(x) + \varepsilon g(x)$ , one can make the parameter  $\varepsilon$  tend to 0. One may naturally expect that the sequence of the resulting minimum points will converge to a solution of the initial problem. This is the essence of the regularization method. The quantity  $\varepsilon$  is called the *regularization parameter* and the function  $g(x)$  is called the *regularization function*.

We examine first the regularization method in the ideal form, where the minimum of the auxiliary problem is exact.

**THEOREM 4.** Let  $f(x)$  be a convex continuous function in  $\mathbb{R}^n$  having a non-empty set of minimum points  $X^*$ , and let  $g(x)$  be a strongly convex continuous function. Let

$$x_\varepsilon = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \Phi_\varepsilon(x), \quad \Phi_\varepsilon = f(x) + \varepsilon g(x), \quad \varepsilon > 0. \quad (11) \quad V(x)$$

Then  $x_\varepsilon \rightarrow x^*$  as  $\varepsilon \rightarrow +0$ , where  $x^*$  is the minimum point of  $f(x)$  for which  $g(x)$  is minimal, i.e.,  $x^* = \underset{x \in X^*}{\operatorname{argmin}} g(x)$ .

**PROOF.** The function  $f + \varepsilon g$  is strongly convex. Hence  $x_\varepsilon$  exists and is unique. Furthermore, from the definition of  $x_\varepsilon$  for an arbitrary  $\hat{x} \in X^*$  we obtain  $f(x_\varepsilon) + \varepsilon g(x_\varepsilon) \leq f(\hat{x}) + \varepsilon g(\hat{x})$ ,  $f(\hat{x}) \leq f(x_\varepsilon)$ , i.e.,  $g(x_\varepsilon) \leq g(\hat{x})$ ,  $g(x_\varepsilon) \leq g(x^*)$ , as well. Since  $g(x)$  is strongly convex, the set  $\{x: g(x) \leq \alpha\}$  is bounded, i.e., the set of  $x_\varepsilon$  is bounded. Take a subsequence  $x_{\varepsilon_i}$  converging to a point  $\tilde{x}$ . Since  $f(x)$  and  $g(x)$  are continuous, then  $\lim_{i \rightarrow \infty} g(x_{\varepsilon_i}) = g(\tilde{x})$ , i.e.,  $g(\tilde{x}) \leq g(x^*)$ , and passing to the limit in the inequality  $f(x_{\varepsilon_i}) + \varepsilon_i g(x_{\varepsilon_i}) \leq f(x^*) + \varepsilon g(x^*)$ , yields  $f(\tilde{x}) \leq f(x^*)$ . Thus,  $\tilde{x} \in X^*$ , and from the inequality  $g(\tilde{x}) \leq g(x^*)$  and the definition of  $x^*$  it then follows that  $\tilde{x} = x^*$ . Thus  $x_{\varepsilon_i} \rightarrow x^*$ . But  $\ell\varepsilon \|x_\varepsilon - x^*\|^2 \leq \varepsilon(g(x^*) - g(x_\varepsilon))$ , i.e., every sequence  $x_\varepsilon$  converges to  $x^*$ .  $\square$

Of course, it is, as a rule, impossible to use the regularization method as described, because the auxiliary problem cannot be solved exactly. One of the few cases where the problem can have, in principle, an exact solution, involves quadratic functions. Let  $f(x) = (Ax, x)/2 - (b, x)$ , where  $A \geq 0$ , and let  $f(x)$  take on a minimum on  $\mathbf{R}^n$  on the nonempty set  $X^*$ . Also, let

$$g(x) = (B(x - a), x - a)/2, \quad (12)$$

where  $B > 0$ . Then in the regularization method the quadratic function is minimized at each step, and therefore

$$x_\varepsilon = (A + \varepsilon B)^{-1}(b + \varepsilon Ba). \quad (13)$$

~~|By (12) the matrix  $A + \varepsilon B$  has an inverse for any  $\varepsilon > 0$ . Theorem 4 implies that  $x_\varepsilon \rightarrow x^* \in X^*$ , where  $x^* = \arg \min_{x \in X^*} g(x)$ . In particular, for  $B = I$ ,  $a = 0$  (i.e., when the regularization function has the form  $g(x) = \|x\|^2/2$ ), then  $x^*$  is the minimum point of  $f(x)$  with smaller norm (it is called the *normal solution* of the problem). In this case we have~~

$$x_\varepsilon = (A + \varepsilon I)^{-1}b. \quad (14)$$

The regularization method for a quadratic problem is closely related to the notion of the so-called *pseudoinverse* matrix. Let  $C$  be an arbitrary  $m \times n$  matrix (not necessarily square). Then the function

$$f(x) = \|Cx - d\|^2, \quad x \in \mathbf{R}^n, \quad (15)$$

attains a minimum on  $\mathbf{R}^n$  (see Exercise 2 of Section 1.3). The minimum point of  $f(x)$  with the smallest norm (the normal solution,  $x^*$ ) is unique. It can be shown that  $x^*$  depends linearly on  $d$ :

$$x^* = C^+d, \quad (16)$$

where  $C^+$  is some  $n \times m$  matrix, referred to as the *pseudoinverse* of  $C$ . It follows from Theorem 4 and equality (14) that

$$C^+ = \lim_{\varepsilon \rightarrow 0} (C^T C + \varepsilon I)^{-1} C^T.$$

Other properties of pseudoinversion are given in Exercise 6 below.

Let us return to the regularization method. Clearly, the smaller  $\varepsilon$ , the closer  $x_\varepsilon$  to a solution, so that taking very small  $\varepsilon$  seems to be appropriate. However, we will see in our later discussion that it cannot be done because of the errors in computing the function and the gradient, as well as the

roundoff errors in solving the auxiliary problem. Then there arises the question whether the solution yielded by the regularization method can be exact for finite  $\varepsilon$ . Here are examples to illustrate that  $\|x_\varepsilon - x^*\|$  can be large even for small  $\varepsilon$ .

Let

$$f(x) = p^{-1}x^p, \quad x \in \mathbf{R}^1, \quad p > 2, \quad g(x) = (x-1)^2/2.$$

Then  $x^* = 0$ , and it is not hard to obtain  $|x_\varepsilon - x^*| = |x_\varepsilon| \approx \varepsilon^{1/(p-1)}$ . Hence, if  $p$  is large, then  $|x_\varepsilon - x^*|$  is relatively large even for small  $\varepsilon$ . Thus, for  $p = 7$ ,  $\varepsilon = 10^{-6}$ , we get  $|x_\varepsilon - x^*| \approx 10^{-1}$ .

2. The Prox - method. The regularization method for the regularization function  $g(x) = \|x-a\|^2/2$  is

$$x^{k+1} = x_{\varepsilon_k} = \operatorname{argmin}_{x \in \mathbf{R}^n} \left( f(x) + \left( \frac{\varepsilon_k}{2} \right) \|x-a\|^2 \right), \quad \varepsilon_k \rightarrow 0.$$

We can try to go the different way: at each step, instead of the regularization parameter  $\varepsilon_k$  we vary the point  $a$ , but replace it with  $x^k$ . We thus arrive at the method

$$x^{k+1} = \operatorname{argmin}_{x \in \mathbf{R}^n} (f(x) + \frac{1}{2} \varepsilon \|x-x^k\|^2), \quad \varepsilon > 0, \quad (18)$$

which is called the *proximal method* (or the *prox-method*), due to its close relation with the so-called proximal mapping. Let  $f(x)$  be a convex function on  $\mathbf{R}^n$  and let  $\varepsilon > 0$  be some parameter. Then the operator

$$\operatorname{Prox} a = \operatorname{argmin}_{x \in \mathbf{R}^n} (f(x) + \frac{1}{2} \varepsilon \|x-a\|^2)$$

is called the *proximal operator*. Its properties and the explicit form for a number of examples are given in Exercises 7 and 8 below.

Now we can write the method as the following:

$$x^{k+1} = \operatorname{Prox} x^k. \quad (19)$$

**THEOREM 5.** Let  $f(x)$  be a convex function on  $\mathbf{R}^n$  with a nonempty set of minimum points  $X^*$ ,  $\varepsilon > 0$ . Then method (19) converges to some point  $x^* \in X^*$ .

**PROOF.** According to Exercise 7, the function

$$\psi(a) = \min_x [f(x) + \frac{1}{2} \varepsilon \|x-a\|^2]$$

*L A* is convex, differentiable,  $\nabla\psi(a) = \varepsilon(a - \text{Prox } a)$  satisfies a Lipschitz condition with constant  $\varepsilon$  and  $X^* = \underset{a}{\text{Argmin}} \psi(a) \neq \emptyset$ . To minimize  $\psi(a)$  we apply the gradient method with  $\gamma = 1/\varepsilon$ :

$$a^{k+1} = a^k - \varepsilon^{-1} \nabla\psi(a^k) = a^k - \varepsilon^{-1} \varepsilon(a^k - \text{Prox } a^k) = \text{Prox } a^k.$$

In other words, the prox-method (19) can be viewed as the gradient method for minimizing  $\psi(a)$ . Applying Theorem 1 (all of its conditions are satisfied) we have what was to be proved. *✓ □*

The advantage of the prox-method versus the regularization method is that the condition number of the auxiliary problems is not affected (the parameter  $\varepsilon$  remains constant). However, the prox-method (just like the gradient method) does not generally lead to a normal solution.

For a quadratic function of the form (6) the prox-method can be written in the explicit form:

$$x^{k+1} = (A + \varepsilon I)^{-1}(b + \varepsilon x^k). \quad (20)$$

3. Iterative regularization. In the foregoing methods, we assumed that at each step an auxiliary problem of unconstrained minimization is being solved (exactly or approximately), without, however, assigning a fixed method for solution. For the case of iterative regularization we, instead, choose some method of unconstrained minimization and execute several iterations for the next auxiliary problem (the number of these iterations can be selected *a priori*, or regulated during the computations). In the simplest variant of such methods one step of gradient descent is made in order to minimize the regularization function, with the regularization parameter changed thereupon. We have thus obtained the method of iterative regularization:

$$x^{k+1} = x^k - \gamma_k(\nabla f(x^k) + \varepsilon_k \nabla g(x^k)), \quad (21)$$

where  $g(x)$  is the regularization function,  $\varepsilon_k$  is the regularization parameter varying in each iteration.

**THEOREM 6.** Let  $f(x)$ ,  $g(x)$  be twice differentiable functions on  $\mathbf{R}^n$ , where

$$\|\nabla^2 f(x)\| \leq L, \quad \ell I \leq \nabla^2 g(x) \leq LI, \quad \ell > 0,$$

$$\text{for all } x, \quad X^* = \underset{x \in \mathbf{R}^n}{\text{Argmin}} f(x) \neq \emptyset,$$

and

$$0 \leq \frac{\varepsilon_{k-1} - \varepsilon_k}{\varepsilon_k^2} \rightarrow 0, \quad 0 \leq \varepsilon_k \rightarrow 0, \quad \sum_{k=0}^{\infty} \varepsilon_k = \infty, \quad (22)$$

$$\gamma_k = \gamma, \quad 0 < \gamma < \frac{2}{(1 + \varepsilon_0)L}. \quad (23)$$

Then in method (21),  $x^k \rightarrow x^*$  where  $x^* \in X^*$ ,  $x^* = \arg \min_{x \in X^*} g(x)$ .

**PROOF.** Let  $y^k \sqrt{\arg \min_{x \in \mathbb{R}^n} \Phi_k(x)}$ ,  $\Phi_k(x) = f(x) + \varepsilon_k g(x)$ ; by the assumptions  $y^k$  exists, is uniquely defined and  $y^k \rightarrow x^*$  (see Theorem 4). The function  $\Phi_k(x)$  is strongly convex with constant  $\ell\varepsilon_k$ . Hence (see (35) of Section 1.1)

$$\Phi_k(y^{k-1}) \geq \Phi_k(y^k) + (\ell\varepsilon_k/2)\|y^k - y^{k-1}\|^2.$$

Similarly, from the strong convexity of  $\Phi_{k-1}(x)$  we get

$$\Phi_{k-1}(y^k) \geq \Phi_{k-1}(y^{k-1}) + (\ell\varepsilon_{k-1}/2)\|y^k - y^{k-1}\|^2.$$

Adding these inequalities yields

$$(\varepsilon_k - \varepsilon_{k-1})(g(y^k) - g(y^{k-1})) + \ell(\varepsilon_{k-1} + \varepsilon_k)\|y^k - y^{k-1}\|^2/2 \leq 0.$$

Since  $\{y^k\}$  is bounded, there is a constant  $M$  such that

$$\|g(y^k) - g(y^{k-1})\| \leq M\|y^k - y^{k-1}\|.$$

Hence

$$\|y^k - y^{k-1}\| \leq \frac{2M(\varepsilon_{k-1} - \varepsilon_k)}{\ell(\varepsilon_{k-1} + \varepsilon_k)} \leq N \frac{\varepsilon_{k-1} - \varepsilon_k}{\varepsilon_k}, \quad N = \frac{M}{\ell}. \quad (24)$$

Now we estimate in method (21) the distance from  $x^{k+1}$  to  $y^k$ :

$$\|x^{k+1} - y^k\| = \|x^k - y^k - \gamma \nabla \Phi_k(x^k)\| = \|x^k - y^k - \gamma A(x^k - y^k)\|.$$

Here, by virtue of (13) of Section 1.1 and the condition  $\nabla \Phi_k(y^k) = 0$ , we have

$$A = \int_0^1 \nabla^2 \Phi_k(y^k + \tau(x^k - y^k)) d\tau.$$

By our assumptions,

$$\ell\varepsilon_k I \leq \nabla^2 \Phi_k(x) \leq L(1 + \varepsilon_k)I \leq L(1 + \varepsilon_0)I.$$

$T \subset$  Hence  $\ell \varepsilon_k I \leq A \leq \bar{f}(1 + \varepsilon)I$  and

$$\begin{aligned} \text{~1/0) } \|x^{k+1} - y^k\| &\leq \|I - \gamma A\| \|x^k - y^k\| \\ &\leq \max_{\ell \varepsilon_k \leq \lambda \leq L(1 + \varepsilon_0)} |1 - \gamma \lambda| \|x^k - y^k\| = (1 - \gamma \ell \varepsilon_k) \|x^k - y^k\| \end{aligned} \quad (25)$$

for sufficiently large  $k$  as  $\varepsilon_k \rightarrow 0$ . Using (24) and (25), we get

$$\begin{aligned} \|x^{k+1} - y^k\| &\leq (1 - \gamma \ell \varepsilon_k) \|x^k - y^k\| \\ &\leq (1 - \gamma \ell \varepsilon_k) \|x^k - y^{k-1}\| + (1 - \gamma \ell \varepsilon_k) \|y^k - y^{k-1}\| \\ &\leq (1 - \gamma \ell \varepsilon_k) \|x^k - y^{k-1}\| + \mu_k, \\ \mu_k &= (1 - \gamma \ell \varepsilon_k) N(\varepsilon_{k-1} - \varepsilon_k) \varepsilon_k^{-1}. \end{aligned}$$

Applying Lemma 3 of Section 2.2 for  $u_k = \|x^k - y^{k-1}\|$  while taking (22) into account yields  $u_k \rightarrow 0$ . But  $\|x^k - x^*\| \leq \|x^k - y^{k-1}\| + \|y^{k-1} - x^*\| \rightarrow 0$  since  $\|x^k - y^{k-1}\| \rightarrow 0$  by what was shown above, and  $\|y^{k-1} - x^*\| \rightarrow 0$  by Theorem 4.  $\square$

With regard to the convergence rate, we see that by the condition  $\sum_{k=0}^{\infty} \varepsilon_k = \infty$ , the parameter  $\varepsilon_k$  cannot tend to zero too rapidly. On the other hand, the method converges not more rapidly than the method of regularization, and the latter, as we saw earlier, may converge slowly.

## Exercises

5. Let  $f(x)$  be a convex function in  $\mathbf{R}^n$ ,  $X^* = \operatorname{Argmin}_{x \in \mathbf{R}^n} f(x) \neq \emptyset$ , let the function  $g(x)$  be strictly convex, and let the set  $\{x: g(x) \leq \alpha\}$  be bounded and nonempty for some  $\alpha$ . Prove (by the same scheme as Theorem 4) the convergence of the regularization method in this case.
6. Using the definition of  $C^+$  and formula (7), prove the following properties of pseudoinverse matrices:
  - a) if  $m = n$  and  $C^{-1}$  exists, then  $C^+ = C^{-1}$ ;
  - b)  $AA^+A = A$ ,  $A^+AA^+ = A^+$ ;
  - c)  $(A^+)^+ = A$ ;
  - d)  $(A^T)^+ = (A^+)^T$ .
7. Prove the following properties of the Prox-operator:
  - a) it is uniquely defined;
  - b) it is nonexpanding, i.e.,  $\|\operatorname{Prox} a - \operatorname{Prox} b\| \leq \|a - b\|$ ;

c) the function

$$\psi(a) = \min_x (f(x) + (\frac{\varepsilon}{2}) \|x-a\|^2)$$

is convex, differentiable, its gradient satisfies a Lipschitz condition with constant  $\varepsilon$  and is equal to  $\nabla\psi(a) = \varepsilon(a - \text{Prox } a)$ ;

d) if

$$X^* = \underset{x}{\operatorname{Argmin}} f(x) \neq \emptyset,$$

then

$$X^* = \underset{a}{\operatorname{Argmin}} \psi(a).$$

8. Compute  $\text{Prox } a$  and  $\psi(a)$  (Exercise 7) for the following examples:

- a)  $f(x) = (Ax, x)/2 - (b, x)$ ,  $A \geq 0$ ;
- b)  $f(x) \equiv 0$ ;
- c)  $f(x) = \|x\|$ .

ANSWERS:

a)  $\text{Prox } a = (A + \varepsilon I)^{-1}(b + \varepsilon a)$ ,

$$\psi(a) = (\frac{1}{2})[\varepsilon \|a\|^2 - ((A + \varepsilon I)^{-1}(b + \varepsilon a), (b + \varepsilon a))]$$

b)  $\text{Prox } a = a$ ,  $\psi(a) \equiv \frac{1}{2}\|a\|^2$

c)  $\text{Prox } a = [1 - \frac{1}{2}/(\varepsilon \|a\|)]_+ a$ ,  $\psi(a) = \varepsilon \|a\|^2/2$  for  $\|a\| \leq \frac{1}{2}/\varepsilon$ ,

$$\psi(a) = \|a\| \text{ for } \|a\| > \frac{1}{2}/\varepsilon.$$

10

71

1 - 1/(2\varepsilon)

### 6.1.3 Methods in the Presence of Noise

Methods for finding a singular minimum have been analyzed above in the idealized situation, when the values of the gradient of the objective function are known exactly (in the gradient method, the conjugate-gradient method, and the method of iterative regularization), or when the auxiliary minimization problem in each iteration is solved exactly (in the regularization method and the prox-method). Let us examine the effect of noise, restricting our analysis to the most typical cases.

1. The gradient method. Suppose there are deterministic errors in calculating the gradient, i.e., at the point  $x^k$  the vector  $s^k$  is admissible:

$$s^k = \nabla f(x^k) + r^k, \quad \|r^k\| \leq \varepsilon. \quad (26)$$

In this situation, the gradient method (1) takes on the form

$$x^{k+1} = x^k - \gamma s^k. \quad (27)$$

As is seen from Theorem 1 of Section 4.2, for a nonsingular minimum one can guarantee convergence into some region around the minimum point  $x^*$ . The radius of this region (see Exercise 2 in Section 4.2) depends on the constant of strong convexity  $\ell$  and tends to infinity as  $\ell \rightarrow 0$ . Hence, it is impossible to draw from these results any inferences concerning the behavior of the method in the singular case (except that the method is ineffective). Actually, on hitting a region of small values of the gradient, method (27) behaves nonsensically—the direction of motion becomes almost arbitrary. Hence method (27) has to be modified, viz. to stop the iterations as soon as the quantity  $\|s^k\|$  becomes sufficiently small. In this form the method turns out to be effective in a certain sense.

**THEOREM 7.** Let  $f(x)$  be a convex differentiable function in  $\mathbb{R}^n$ , the gradient of which satisfies a Lipschitz condition with constant  $L$ , and let  $X^* = \underset{x \in \mathbb{R}^n}{\operatorname{Argmin}} f(x) \neq \emptyset$ . Suppose the quantities  $L, \varepsilon$  are known (see (26)) and  $\rho \geq \|x^0 - x^*\|$ , where  $x^* = P_{X^*}(x^0)$  is the minimum point closest to  $x^0$ . Let the iterations (27) with  $0 < \gamma < 2/L$  be continued until the condition

$$\|s^k\| \leq \varepsilon + 2\sqrt{\frac{\varepsilon L \rho}{2 - L\gamma}} \quad (28)$$

is satisfied, and let  $x_\varepsilon$  be the point of  $x^k$  at which this condition is satisfied for the first time. Then the process stops in not more than  $\rho/(\gamma\varepsilon) + 1$  iterations, and also

$$\|\nabla f(x_\varepsilon)\| \leq \left( \varepsilon + \sqrt{\frac{\varepsilon L \rho}{2 - L\gamma}} \right) \quad \text{and} \quad \|x_\varepsilon - x^*\| \leq \rho .$$

**PROOF.** From (26) and (27) we have

$$\|x^{k+1} - x^*\| = \|x^k - x^* - \gamma \nabla f(x^k) - \gamma r^k\| \leq \|x^k - x^* - \gamma \nabla f(x^k)\| + \gamma \varepsilon_k .$$

From inequality (3) (with  $\hat{x}$  replaced by  $x^*$ ), we have

$$\|x^k - x^* - \gamma \nabla f(x^k)\|^2 \leq \|x^k - x^*\|^2 - \gamma(2/L - \gamma) \|\nabla f(x^k)\|^2 .$$

For arbitrary  $a \geq b > 0$  we have the inequality  $\sqrt{a^2 - b^2} \leq a - b^2/(2a)$ , so that

$$\|x^k - x^* - \gamma \nabla f(x^k)\| \leq \|x^k - x^*\| - \gamma(2 - L\gamma) \|\nabla f(x^k)\|^2 (2L \|x^k - x^*\|)^{-1} .$$

Thus

$$\|x^{k+1} - x^*\| \leq \|x^k - x^*\| - \frac{\gamma(2 - L\gamma) \|\nabla f(x^k)\|^2}{2L \|x^k - x^*\|} + \gamma \varepsilon . \quad (29)$$

Suppose that  $x^k$  is not the stopping point. Then

$$\varepsilon + 2\sqrt{\frac{\varepsilon L\rho}{2 - L\gamma}} \leq \|s^k\| \leq \|\nabla f(x^k)\| + \varepsilon$$

yielding

$$\|\nabla f(x^k)\|^2 \geq (4\varepsilon L\rho)/(2 - L\gamma).$$

Substituting this into (29), we get

$$\|x^{k+1} - x^*\| \leq \|x^k - x^*\| - \gamma\varepsilon \left( \frac{2\rho}{\|x^k - x^*\|} - 1 \right). \quad (30)$$

Since  $\|x^0 - x^*\| \leq \rho$ , one has  $\|x_1 - x^*\| \leq \rho - \gamma\varepsilon$  and generally  $\|x^k - x^*\| \leq \rho - k\gamma\varepsilon$  for all  $k$  until the process stops. Hence the number of iterations before the process stops does not exceed  $\rho/(\gamma\varepsilon) + 1$ . Since at the stopping point

$$\|\nabla f(x^k)\| - \varepsilon \leq \|s^k\| \leq \varepsilon + 2\sqrt{\varepsilon L\rho/(2 - L\gamma)},$$

then

$$\|\nabla f(x_\varepsilon)\| \leq (\varepsilon + \sqrt{\varepsilon L\rho/(2 - L\gamma)}). \quad \square$$

Let us examine this result more closely. In the modification of the gradient method, it is guaranteed that (1) a point will be obtained with sufficiently small norm of the gradient:  $\|\nabla f(x_\varepsilon)\| \leq \phi(\varepsilon)$ , where  $\phi(\varepsilon) = O(\sqrt{\varepsilon}) \rightarrow 0$  as  $\varepsilon \rightarrow 0$  and (2) this point is not too far away from the minimum point closest to the initial approximation. Using the inequality  $f(x_\varepsilon) - f(x^*) \leq \|\nabla f(x_\varepsilon)\| \|x_\varepsilon - x^*\|$  one can guarantee that at  $x_\varepsilon$  the value of the function is also close to the minimal value:

$$f(x_\varepsilon) - f(x^*) = O(\sqrt{\varepsilon}). \quad (31)$$

In this sense, the point  $x_\varepsilon$  is expected to yield an approximate solution of the minimization problem. Of course, it is impossible to give any explicit bound of how close  $x_\varepsilon$  is to  $x^*$ .

For problems with random noise one can prove a result on almost sure convergence of the gradient method, in which the step size  $\gamma_k$  tends to zero (see Exercise 9).

If the noise level depends on the number of the iteration:  $\|r^k\| \leq \varepsilon_k$ , then for  $\sum_{k=0}^{\infty} \varepsilon_k < \infty$  the gradient method converges in the usual sense.

2. The regularization method. Due to the inevitable errors in calculating  $f(x)$ , as well as the impossibility of finding the

exact minimum of a nonquadratic function, the auxiliary problem of unconstrained minimization (11) in the regularization method can be solved only approximately, with accuracy up to some quantity  $\delta$ . Let

$$\Phi_\varepsilon(x_\varepsilon^\delta) \leq \Phi_\varepsilon^* + \delta, \quad (32)$$

where

$$\Phi_\varepsilon(x) = f(x) + \varepsilon g(x), \quad \Phi_\varepsilon^*(x) = \min_{x \in \mathbb{R}^n} \Phi_\varepsilon(x).$$

**THEOREM 8.** Let the conditions of Theorem 4 be satisfied. Then as

$$\varepsilon \rightarrow 0, \quad \delta/\varepsilon \rightarrow 0 \quad (33)$$

one has  $x_\varepsilon^\delta \rightarrow x^*$ .

**PROOF.** Is the same as that of Theorem 4.  $\square$

It is impossible to give bounds of the closeness  $\|x_\varepsilon^\delta - x^*\|$  in explicit form for an arbitrary function  $f(x)$  (see the examples related to Theorem 4).

**3. Other methods.** The other methods described earlier can be treated in similar fashion, in particular, the prox-method and the method of iterative regularization. We shall not dwell on this at length since both the technique and the results are similar to Theorems 7 and 8.

### Exercise

**9.** Let  $f(x)$  be a convex differentiable function in  $\mathbb{R}^n$ , where  $\nabla f(x)$  satisfies a Lipschitz condition,  $X^* = \underset{x \in \mathbb{R}^n}{\operatorname{Argm}\nolimits} f(x) \neq \emptyset$ . Let  $s^k = \nabla f(x^k) + \xi^k$ , where the random noise  $\xi^k$  is independent and  $E\xi^k = 0$ ,  $E\|\xi^k\|^2 \leq \sigma^2$ . Consider the gradient method  $x^{k+1} = x^k - \gamma_k s^k$  under the conditions  $\sum_{k=0}^{\infty} \gamma_k = \infty$ ,  $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$ . Using the method of proof of Theorem 1 of this section and Theorem 1 of Section 2.2, prove that  $x^k \rightarrow \bar{x} \in X^*$  a.s., where the point  $\bar{x}$  can vary for varied realizations of the process.

#### 6.1.4 Summary

Now the time has come to answer the major question, Can one solve optimization problems with singular minimum in actual practice? The answer to this question is not so simple as one may think, and makes one review again the relationship of the theoretical results on convergence and the

practical calculations, which we discussed in Section 1.6. The fact is that mathematicians regard this question as inappropriate and, instead, limit themselves to the results of the type described. However, it is not clear at all what Theorem 8 may lead to; and we leave it to the interested reader to ponder this problem as an exercise.

Of primary importance is the understanding of what exactly is required of an approximate solution to an optimization problem. This depends on the further use of this solution. In some cases, we are primarily interested in determining the minimum point (these are the so-called argument minimization problems). For example, estimation of physical constants on the basis of both direct and indirect measurements is reduced (by the maximum likelihood method or by the least squares method, see Chapter 11) to minimizing a particular function. In that case, the actual arguments of the minimum of this function have a direct physical meaning, and the values deduced, that is the estimates of the sought parameters, will be used in various problems unrelated to the initial minimization problem. Hence it is crucial to find the minimum point as accurately as possible, i.e., we have an argument problem. The situation is similar in some other problems of estimation and identification. However, in the majority of cases, the coordinates of the minimum point are of no consequence; to guarantee the smallest possible value of an optimality criterion is most important. These are criterion optimization problems. Say, in best approximation problems it is required to approximate a given function  $a(t)$  by some simpler expression, e.g. a polynomial of degree  $n$ ,  $\sum_{i=1}^{n+1} x_i t^{i-1}$ . After the appropriate norm ( $L_1$ ,  $L_2$ ,  $L_\infty$  and such) is chosen, the problem reduces to minimizing the function  $f(x) = \|a(t) - \sum_{i=1}^{n+1} x_i t^{i-1}\|$ . However, the values of the coefficients  $x_i^*$  minimizing the  $f(x)$  are of no importance; it is the smallness of the  $f(x)$  that is important. Furthermore, instead of algebraic polynomials we could choose trigonometric polynomials, or seek an approximation over some other class of functions. A very similar situation occurs in many other problems in which a system has to be described optimally by means of a model, the choice of which is arbitrary to some degree, while the goal is to minimize the “discrepancy” between the outputs of the model and of the system. Optimization problems in economics, or problems of optimal design are other examples of criterion problems.

In tackling criterion problems, a singularity of the minimum presents no difficulty since we need only to get into a region of small values of the objective function  $f(x)$ . A formal proof of this assertion is provided by the bounds (2), (10), (31) of the accuracy of approximate solutions obtained by varied methods. Thus, the bound (10) shows that in minimizing without noise an arbitrary quadratic function (possibly, with singular minimum) regardless the dimension of the space, the conjugate-gradient method guarantees the bound  $f(x^k) - f^* = O(k^{-2})$ . This means that in

100 iterations the value of the function can be diminished by a factor of 10,000, which is usually sufficient for practical purposes. In the presence of noise, the bound (31) guarantees that if the noise level is low, the gradient method with the stopping rule (28) makes it possible to find a sufficiently good approximation with respect to the function, regardless of the singularity of the minimum or the dimension of the space. To sum up, it is possible to construct operating algorithms for criterion problems.

The case of argument problems is much more complicated. Note that even in the absence of noise we had results on the convergence of the methods (Theorems 1-6); but we never obtained a bound on convergence rate. As we mentioned earlier, convergence theorems without convergence rate bounds are not an adequate measure of the efficiency of the method. Moreover, the above examples illustrate that the convergence rate of each method discussed can be very small. Hence, none of the above methods guarantees that a singular minimum can be found with a prescribed accuracy (with respect to the arguments) in a number of iterations *a priori* determined. In implementation problems, the computation is complicated by inevitable errors. Results on the behavior of the methods in the presence of noise (Theorems 7, 8) do not contain any bounds of the closeness of the approximate solution to the exact value (with respect to the argument). Theorem 8 provides an asymptotic result: if the noise level tends to 0, then the approximate solutions converge to the exact solution. However, in practice, we are solving a problem for a fixed noise level, and this asymptotic result (without accuracy bounds) provides no information concerning the guaranteed accuracy of the solution.

Our pessimistic approach does not imply, however, that argument singular problems are unsolvable in general. An abundant *a priori* information about a solution is frequently available, and it can be put to use effectively. Thus, if the closeness of the solution to some point  $a$  is known, the latter can be chosen as an initial approximation for the iterative methods (e.g., the gradient method). By Theorem 7, the gradient method guarantees the finding of an approximate solution  $x_\epsilon$  such that  $\|x_\epsilon - x^*\| \leq \|a - x^*\|$ , and  $f(x)$  is the smaller, the smaller the noise level. Another way of taking into account *a priori* information in this case involves a choice of a regularization function of the form  $\|x - a\|^2$ . The available *a priori* information about some properties of the solution can be used in the iterative methods by choosing a suitable norm, and in the regularization method by choosing a particular  $g(x)$ . Furthermore, in statistical problems, such as parameter estimation problems, the information about a solution is usually interpreted in terms of an *a priori* distribution. Taking the Bayesian approach, it is possible to include this information in the objective function, thus helping finding the solution.

To summarize, the possibility of solving argument problems with singular minimum is usually determined by the available *a priori* information

about the solution. Without this information, it is hard to count on obtaining an accurate solution of any kind.

## 6.2 MULTIMODALITY

So far we have basically tackled the problems of minimizing convex functions for which every local minimum coincides with the global one (Theorem 2 of Section 1.2). When the function is multimodal (i.e., it has many local minima), the problem of finding the global minimum is very complex. Throughout this section we shall consider the problem

$$\min f(x) , \quad x \in \mathbf{R}^n , \quad (1)$$

where the function  $f(x)$  is smooth but not convex.

### 6.2.1 Preliminary Remarks

As is known (Theorem 1 of Section 1.2), every local minimum point  $x^*$  in problem (1) is stationary, i.e.,  $\nabla f(x^*) = 0$ . Conversely, if at a stationary point one has  $\nabla^2 f(x^*) > 0$ , then  $x^*$  is a local (or global) minimum point (Theorem 4 of Section 1.2). Similarly, if  $\nabla^2 f(x^*) = 0$  and  $\nabla^2 f(x^*) \neq 0$ , then  $x^*$  is a local maximum point. Finally, if the matrix  $\nabla^2 f(x^*)$  is indefinite at a stationary point  $x^*$ , then we can find vectors  $y$  for which  $f(x^* + \varepsilon y) > f(x^*)$  for sufficiently small  $\varepsilon > 0$ , referred to as directions of increase, as well as vectors  $y$  for which  $f(x^* + \varepsilon y) < f(x^*)$ , referred to as directions of decrease; the point  $x^*$  is called a *saddle point*. Let us state these results as the following theorem.

**THEOREM 1.** Let  $\nabla f(x^*) = 0$ , let the matrix  $\nabla^2 f(x^*)$  be nonsingular, with  $\lambda_1 \leq \dots \leq \lambda_n$  being its eigenvalues and  $e^1, \dots, e^n$  being the corresponding orthonormal eigenvectors. If  $\lambda_1 \neq 0$ , then  $x^*$  is a minimum point; if  $\lambda_n < 0$ , then  $x^*$  is a maximum point; and if  $\lambda_1 < 0 < \lambda_n$ , then  $x^*$  is a saddle point. The vectors  $y \in L_- = \{\sum_{i:\lambda_i < 0} \gamma_i e^i\}$ ,  $y \neq 0$ , are directions of decrease and the vectors  $y \in L_+ = \{\sum_{i:\lambda_i > 0} \gamma_i e^i\}$ ,  $y \neq 0$ , are directions of increase. Here  $\mathbf{R}^n = L_- \oplus L_+$ , i.e.,  $\mathbf{R}^n$  is the direct sum of the subspaces  $L_-$  and  $L_+$ .  $\square$

The point  $x^*$  with  $\nabla f(x^*) = 0$  and nonsingular Hessian is called a *nonsingular stationary point*, the dimension of the subspace  $L_-$  is called the *index of the stationary point*, so that the index is zero if and only if  $x^*$  is a minimum point.

We now turn to analyze the behavior of the major minimization methods in a neighborhood of various stationary points. Let us start with the gradient method of the form  $x^{k+1} = x^k - \gamma \nabla f(x^k)$ . We know (Theorem 4

of Section 1.4) that in a neighborhood of a nonsingular minimum the gradient method for  $0 < \gamma < 2/\|\nabla^2 f(x^*)\|$  converges to  $x^*$ , regardless whether it is a global or a local minimum. Next, let  $x^*$  be a nonsingular stationary point with nonzero index. Then

$$\begin{aligned} x^{k+1} - x^* &= x^k - x^* - \gamma \nabla f(x^k) \\ &= (I - \gamma \nabla^2 f(x^*))(x^k - x^*) + o(x^k - x^*). \end{aligned} \quad (2)$$

If  $x^*$  is a maximum point, the eigenvalues of the matrix  $I - \gamma \nabla^2 f(x^*)$  are greater than 1 for any  $\gamma > 0$  (they are equal to  $1 - \gamma \lambda_i$ ,  $i = 1, \dots, n$ , but all  $\lambda_i \geq 0$ , see Theorem 1). Hence  $\|(I - \gamma \nabla^2 f(x^*))z\| \geq q\|z\|$ ,  $q > 1$ , for all  $z$ . It follows that for sufficiently small  $\|x^k - x^*\| \neq 0$  one will have  $\|x^{k+1} - x^*\| > \|x^k - x^*\|$ . Thus, if  $x^0$  is close to  $x^*$  but does not coincide with  $x^*$ , then iterations in the gradient method go away from the point  $x^*$ . In other words, a maximum point is a point of repulsion for the gradient process, and a trajectory that has hit a neighborhood of such a point will automatically leave this neighborhood (except for the special case where  $x^0 = x^*$ ). L<

For the case where  $x^*$  is a saddle point, the analysis is simple if  $f(x)$  is a quadratic function. Then

$$x^{k+1} - x^* = (I - \gamma A)(x^k - x^*), \quad x^k - x^* = (I - \gamma A)^k(x^0 - x^*), \quad (3)$$

$A = \nabla^2 f(x)$ . In the notation of Theorem 1,  $(x^k - x^*, e^i) = (1 - \gamma \lambda_i)^k (x^0 - x^*, e^i)$ . If  $\lambda_i > 0$ ,  $0 < \gamma < 2/\|A\|$ , then  $|1 - \gamma \lambda_i| < 1$ , and so  $(x^k - x^*, e^i) \rightarrow 0$ . But if  $\lambda_i < 0$ , then  $(x^k - x^*, e^i) = q_i^k (x^0 - x^*, e^i)$ , where  $q_i = 1 - \gamma \lambda_i > 1$ , and hence  $(x^k - x^*, e^i) \rightarrow \infty$  for  $(x^0 - x^*, e^i) \notin 0$ . Since  $\|x^k - x^*\|^2 = \sum_{i=1}^n (x^k - x^*, e^i)^2$ , we get  $\|x^k - x^*\| \rightarrow \infty$  if  $x^0 - x^* \notin L_+$ . Thus, if the initial approximation does not belong to the subspace  $L_+$ , the trajectory of the gradient method will move away from the saddle point. For the nonquadratic case the analysis is more complicated; however, it yields a similar result, i.e., only for an exceptional set of initial points do the gradient iterations lead to a saddle point. L≠

Roughly, the gradient method “almost never” converges to a maximum point or to a saddle point. At the same time, it does not discriminate between a local and a global minimum, and converges arbitrarily to either one.

Newton's method behaves somewhat differently. It follows from Theorem 3 of Section 1.5 that the nonsingularity of  $\nabla^2 f(x^*)$  is sufficient for the method to converge, and  $\nabla^2 f(x^*)$  does not have to be positive definite. Hence Newton's method can converge to any stationary point since it does not distinguish maxima from minima, or from saddle points.

We skip the other minimization methods considered in the preceding chapters. Some of those methods substantially rely on the assumption that

the function is convex, and when this assumption is not satisfied they are no longer effective (almost all the methods given in Chapter 5 are of this kind). Other methods converge to any stationary point (certain variants of quasi-Newton methods). Finally, a larger class of methods converge, as a rule, to an arbitrary local minimum point. It is worth emphasizing that no method guarantees that it hits the global minimum.

### 6.2.2 Exact Methods

All methods of multimodal optimization can be divided into (1) exact methods and (2) heuristic methods. For the exact methods there exist exact assertions concerning their convergence to a global minimum. For the heuristic methods, one has to restrict oneself to some plausible arguments about their rational behavior in the multimodal situation. Exact methods are generally of little value. To illustrate our point, we give a typical example.

**THEOREM 2.** Let  $f(x)$  be a continuous function on the set  $Q = \{a \leq x \leq b\} \subset \mathbf{R}^n$ , and let  $x^k$  be the sequence of independent uniformly distributed random vectors on  $Q$ . Then

$$\min_{1 \leq i \leq k} f(x^i) \xrightarrow{P} \min_{x \in Q} f(x).$$

**PROOF.** By the Weierstrass theorem (Section 1.3) there exists a global minimum point  $x^*$  of  $f(x)$  on  $Q$ . Let  $\varepsilon > 0$  be arbitrary. By the continuity of  $f(x)$  we can find a neighborhood  $U$  of  $x^*$  for which  $f(x) \leq f(x^*) + \varepsilon$  for  $x \in U$ . Let  $v$  denote the volume of  $U \cap Q$  and let  $V$  denote the volume of  $Q$ . Then  $v > 0$  since  $U$  is open. The probability of  $x^k$  being in  $U \cap Q$  is  $v/V$ ; the probability that at least one of the points  $x^1, \dots, x^k$  is in  $U \cap Q$  is  $p_k = 1 - (1 - v/V)^k$ . Obviously,  $p_k \rightarrow 1$  as  $k \rightarrow \infty$ , i.e.,

$$P\left\{\min_{1 \leq i \leq k} f(x^i) > f(x^*) + \varepsilon\right\} \rightarrow 0 \quad \text{as } k \rightarrow \infty.$$

This implies the convergence in probability.  $\square$

Theorem 2 is simple and general, but trivial. Let us take an example and estimate the number of calculations of the function needed in order to find the solution with a small accuracy. Suppose  $x = (x_1, \dots, x_{10}) \in \mathbf{R}^{10}$ ,  $f(x) = \max_{1 \leq j \leq 10} x_j$ ,  $Q = \{x: 0 \leq x_j \leq 1, j = 1, \dots, 10\}$ , with the accuracy at  $\varepsilon = 10^{-2}$ . Then  $x^* = 0, f(x^*) = 0, v = (10^{-2})^{10} = 10^{-20}, V = 1, p_k \approx k \cdot 10^{-20}$ , i.e., for the probability of finding  $x^*$  to within 1% accuracy to be at least equal to 10% one needs to have  $10^{19}$  iterations. In other words, the method of random search (in the form described in Theorem 2) is absolutely useless

for finding the global minimum even for dimensions of order 10. Again we face the fact that the convergence theorem per se does not guarantee the effectiveness of the method. Nevertheless, new works are published now and then, which contain results and a mathematical justification of the methods at the level equal to Theorem 2 (we recall here Wolfe's parody [1.11] in which he discusses a deterministic variant of Theorem 2 in dead earnest).

At the same time, the reader should understand that a method better than Theorem 2 is practically unfeasible for arbitrary continuous, or even smooth functions. Figure 26 illustrates functions for which the global minimum can be found only by selecting its values on a sufficiently fine mesh. Hence we need to restrict the class of the functions. We shall require the functions satisfy the Lipschitz condition

$$|f(x) - f(y)| \leq L \|x - y\|, \quad (4)$$

and assume that the constant  $L$  is known. In minimizing such functions one can be guided by the following considerations. Suppose we have found the best value of  $f(x)$  over the  $k-1$  previous iterations:  $\phi_{k-1} = \min_{1 \leq i \leq k-1} f(x^i)$

and computed the  $f(x^k)$ . Then, if  $f(x^k) < \phi_{k-1}$ , then the best value is improved:  $\phi_k = f(x^k)$ , but if  $f(x^k) > \phi_{k-1}$ , then automatically in the ball  $\{x: \|x - x^k\| < L^{-1}(f(x^k) - \phi_{k-1})\}$  there is no global minimum point, which leads to a reduction of the region of possible localization of the minimum. It is not hard to implement this idea in the computational algorithm for the one-dimensional case. For the multidimensional case, this is difficult to do, because it is not easy to describe the region of localization of the minimum and realize the rule for choosing the new point. How effective such methods are depends on the form of the function as well as on the arrangement of points.

For example, if the difference between  $f(x^1)$  and  $f(x^2)$  is not great, it is possible to cut off at once a region of large volume. If the function is as shown in Figure 26, the method is not superior to the one of complete trials for any rule for choosing  $x^k$ . Furthermore, in practical problems the constant  $L$  in (4) is rarely known, and incorrectly specified value of  $L$  may either slow down the method drastically or lead to a loss of the global minimum.

A similar situation arises if the bounds of the derivatives of the objective function are known. For instance, if  $\nabla f(x)$  satisfies the Lipschitz condition:

$$\|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\| \quad (5)$$

and  $L$  is known, the region can be diminished by means of the inequality

$$|f(x) - f(x^k) - (\nabla f(x^k), x - x^k)| \leq (L/2) \|x - x^k\|^2. \quad (6)$$

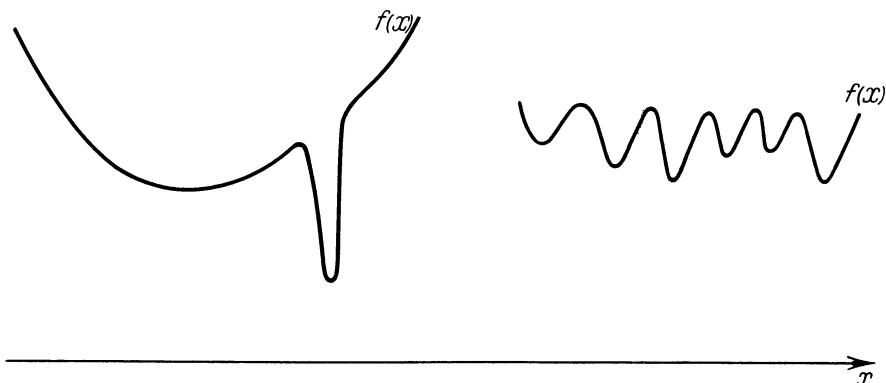


Fig. 26 Functions for which the global minimum is difficult to find.

We limit further discussion since this method has the same drawbacks.

At the present time, there are no other classes of multimodal functions which are natural and easy to describe. On the whole, one can say that the existent exact methods for finding a global extremum cannot be viewed as effective methods for solving multidimensional problems.

### 6.2.3 Deterministic Heuristic Methods

One of the possible approaches to solving multimodal problems is to combine methods of local optimization with a particular procedure of trials for the initial points. For example, one can execute the descent by the conjugate gradient method from the vertices of a coarse uniform grid covering the region of *a priori* localization of the minimum. The initial “trial points” can be laid out differently. Thus, there exist methods for distributing the points “more uniformly” in a multidimensional parallelepiped than at the vertices of a rectangular grid, such as the so-called LP-sequences in [6.14]. Here the number of trial points can be small (a few dozen). The process of a subsequent local minimization is to be stopped if we either hit a zone of local minimum already explored or if the value of the function at a rough local minimum is noticeably greater than the current best value.

Of greater interest are the methods in which the global search is represented as a unified iterative process. In this case, the algorithm needs to be able to “escape” from local minima. The heavy-ball method is the simplest example (Section 3.2), where the approximations \$x^k\$ are related through

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}). \quad (7)$$

Clearly, if  $\nabla f(x^k) = 0$ , but  $x^k \neq x^{k-1}$ , then  $x^{k+1} \neq x^k$ , i.e., the method does not jam at a stationary point. By the mechanical analogy (7), viz. the motion of a heavy ball along the uneven surface, it follows that if the velocity of the ball is sufficiently large, it “skips” over shallow holes. It is possible to show by examples that this method does indeed have the property that it escapes the shallow local minima. However, it may “fall” into a deep minimum and would not be able to get out. Hence the heavy-ball method (7) is not a reliable way to find the global minimum.

The “gully” method suggested by I.M. Gel'fand and M.L. Tsetlin is more promising. This method is based on the gully-shaped objective function, i.e., it is assumed that the function varies weakly in some directions (forming the bottom of the gully) and varies sharply in other directions (the directions of the slopes of the gully). An example of a monomodal gully function is a quadratic function with ill-conditioned matrix. Generally, in a neighborhood of a local minimum gully functions are characterized by a large condition number  $\mu$  (see Section 1.3). The gully method consists of descent steps made by any local method (usually the gradient method) and descending to the bottom of the gully, and of gully steps along the bottom of the gully. The structure of the method is shown in Figure 27, where  $x^0$  and  $x^1$  are two initial approximations, the thin lines denote the

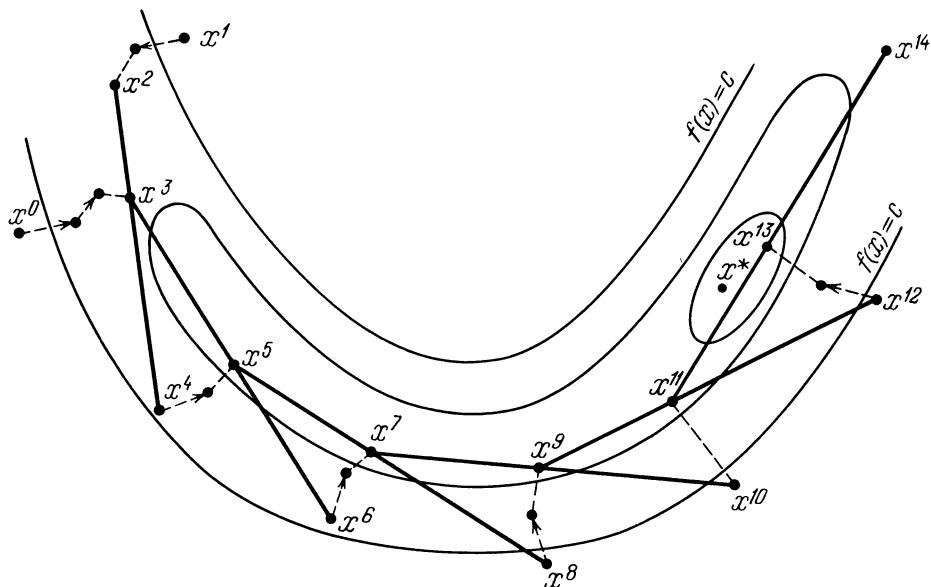
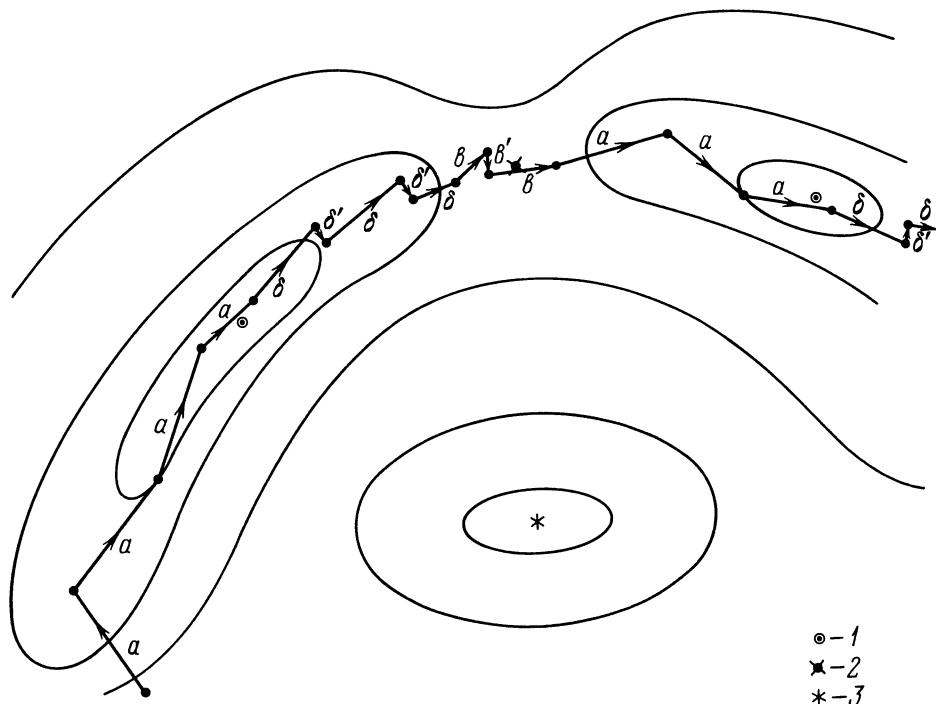


Fig. 27 The gully method.

descent steps, the bold lines show the gully steps. A trajectory in the gully method passes mainly along the bottom of the gully without “sticking” to the local minima (importantly, the gully steps are of a particular size, regardless whether the function increases or decreases in a given direction). The gully method is intended for a cursory inspection of the domain of definition of the function. The points with small values of  $f(x)$  have to be further refined by means of more powerful local methods. Nevertheless, the gully method is not without drawbacks. For example, it is difficult to choose the appropriate size of the gully step—for a large step the method skips over many minima, for a small step the method does not track the bottom of the gully and its motion becomes chaotic. Also, the direction of the gully step is not defined uniquely and depends on many factors, the accuracy of the local descents, the position of the previous point, among others. In general, the presence of many “free” parameters in the gully method explains why the use of this method requires a lot of experience and a thorough preliminary “adjustment.”

The concepts of the gully method are used in the DAS method (descent-ascent saddle method). In this method, the entire procedure of finding the global minimum is divided into three stages, which are repeated cyclically. At the descent stage, the local minimum is found by the conjugate gradient method. At the ascent stage, the method leaves the zone of the minimum. The method moves in the direction of the slowest ascent, which is found in the following way. At a point  $x^k$  the function  $f_k(x) = f(x) - (\nabla f(x^k), x)$  is formed. Obviously,  $\nabla f_k(x^k) = 0$ , and if  $\nabla^2 f(x^k) > 0$ , then  $x^k$  is a local minimum point of  $f_k(x)$ . But if  $\nabla^2 f(x^k)$  is indefinite, then  $x^k$  is a saddle point of  $f_k(x)$ . From the point  $z^0 = x^k + \varepsilon d^{k-1}$  (where the  $d^{k-1}$  is the direction of the previous motion,  $\varepsilon > 0$  is a parameter) several steps of the gradient method are made for  $f_k(x)$ :  $z^{i+1} = z^i - \gamma \nabla f_k(z^i)$ . The fact that the points  $z^i$  tend to  $x^k$  signifies that ~~the~~  $\nabla^2 f_k(x^k) = \nabla^2 f(x^k)$  is positive definite (see the investigation of the behavior of the gradient minimum in Section 6.2.1<sup>7</sup>) in a neighborhood of a local minimum and of a saddle point, and the direction  $d^k = (z^i - x^k)/\|z^i - x^k\|$  is taken as the direction of ascent. It is easy to verify that this direction is close to the eigenvector of  $\nabla^2 f(x^k)$  corresponding to the smallest eigenvalue of the matrix. The step  $\bar{x}^{k+1} = x^k + \lambda_k d^k$  is made, and the gradient step from  $\bar{x}^{k+1}$  leads to a new point  $x^{k+1}$ , at which the ascent procedure is repeated. However, if the points  $z^i$  move away from  $x^k$ , it is clear that  $f(x)$  is not convex in a neighborhood of  $x^k$  and the method moves to the saddle point. The vector  $d^k = (z^i - x^k)/\|z^i - x^k\|$  defines the direction of motion to the saddle point, in combination with the gradient descent after each step. After passing the saddle point (identified by the change of sign of the  $(\nabla f(x^{k+1}), d^k)$ ) the method begins its descent to a new local minimum. Figure 28 shows a typical trajectory of the DAS method. The search in several stages seems to be superior to the unified motion in the gully method.



*L<sub>b</sub>*  
1-<sub>stage</sub>

Fig. 28 The DAS method for global minimization:  $a$  is the descent stage;  $\delta$  is the ascent stage;  $\star$  is the saddle point; 1 denotes minimum points; 2 denotes the saddle point; 3 denotes the maximum point.

There are many other heuristic methods of global optimization. Unfortunately, no strict results on their effectiveness have been obtained so far, and their verification using test problems is not always convincing, nor sufficiently thorough.

#### 6.2.4 Stochastic Heuristic Methods

Two approaches can be distinguished in this respect, involving (1) a randomness in the minimization process (the method of random search) and (2) a stochastic model of the objective function.

Methods of random search for local optimization were described in Section 3.4. To make these methods of global nature, it is required to allow large steps to lead the method from the neighborhood of a local minimum. Here is the simplest variant of such a method. Suppose one seeks the global minimum of  $f(x)$ ,  $x \in \mathbf{R}^n$ , on the unit cube  $Q = \{x: 0 \leq x_i \leq 1\}$ . At the point  $x^k$  one chooses the vector  $h^k$  with independent components

uniformly distributed on  $[-1, 1]$ , and if  $x^k + h^k \in Q$  and  $f(x^k - h^k) < f(x^k)$ , then one takes  $x^{k+1} = x^k + h^k$ . Otherwise, one takes a new realization of  $h^k$ . The method seems to be quite well-founded, a theorem on convergence can be proved for this method, etc. However, it is easy to see that it coincides (within the notation) with the method of Theorem 2, i.e., it is completely ineffective, as we showed before. Unfortunately, the same danger awaits us, too, in other methods of random search, although it may not be so obvious as in our “naïve” variant of the method. Hence it is hard to share the optimistic enthusiasm of the random search advocates, who seem to believe that they indeed possess an effective tool of global minimization. The reader who wishes to learn more about varied modifications of random search methods is advised to read the extensive existent literature in the subject matter.

The other approach involving randomness in global optimization is based on the idea that upon calculation of the objective function at  $k$  points  $x^1, \dots, x^k$ , it is possible to speak of the probabilities of its values at the remaining points. In this case, the notion of “probability” is given sometimes a precise meaning: it is assumed that there is a class of functions with probability measure defined on it, the objective function  $f(x)$  belonging to this class. Then it is possible to speak of conditional probabilities of various events under the realization of the values  $f(x^1), \dots, f(x^k)$ . Most often, however, a nonstrict probability model is used. Usually one assumes that a realization of the value of  $f(x^k)$  at  $x^k$  “enhances the probability” of values of  $f(x)$  close to  $f(x^k)$  for points in a neighborhood of  $x^k$  and does not change them far away from the  $x^k$ . With a specified *a priori* distribution and a sufficiently arbitrary rule of updating, *a posteriori* distribution of the values of  $f(x)$  for all  $x$  are obtained. The point at which the “mathematical expectation” of  $f(x)$  is minimal is taken as the point  $x^{k+1}$ , and after the  $f(x^{k+1})$  has been calculated the probabilities are computed again. Many concrete implementations of this idea are known at the present time. For all similar methods it is difficult to describe the *a posteriori* probabilities as well as procedures for finding the “best” points. Moreover, the basis for methods of this kind is not well-founded.

We have made an attempt to present the state of the art in the area of global minimization. As the reader can see, the situation is far from perfect. Further in-depth studies, theoretical as well as numerical, of the existent methods are needed. New ideas are necessary, most of all in the classification of multimodal problems and in the determination of relatively narrow classes of problems, permitting special, efficient methods for their solution.

### 6.3 NONSTATIONARY PROBLEMS

In some engineering problems involving on-line control of systems, the optimality criterion does not remain constant but, rather, is time variant (e.g., due to the drift of the system's characteristics). The accuracy in stating such nonstationary problems of optimization depends on the control and the information available to the user. We shall now briefly describe some possible situations.

#### 6.3.1 The Form of $f(x, t)$ is Known

Suppose that the objective function depends on a scalar parameter  $t$  (not necessarily time), i.e., it has the form  $f(x, t)$ ,  $x \in \mathbb{R}^n$ ,  $t \in \mathbb{R}^1$ . We write the local or global minimum of  $f(x, t)$  for fixed  $t = t_0$  as

$$x_0^* = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f(x, t_0). \quad (1)$$

Then, by the necessary conditions for a minimum,  $x_0^*$  is a solution of the equation

$$\nabla_x f(x, t_0) = 0.$$

If we assume that the sufficient conditions for a minimum  $\nabla_{xx}^2 f(x_0^*, t_0) > 0$ , is satisfied, the matrix  $\nabla_{xx}^2 f(x, t)$  is continuous at  $\{x_0^*, t_0\}$ ,  $\nabla_x f(x, t)$  is differentiable in  $t$  at  $\{x_0^*, t_0\}$  and  $\nabla_x f(x, t)$  is continuous in a neighborhood of  $\{x_0^*, t_0\}$ , then the conditions of the implicit function theorem are satisfied (Theorem 2 of Section 2.3), and hence in a neighborhood of  $t_0$  there exists a differentiable function  $x^*(t)$  for which  $\nabla_x f(x^*(t), t) = 0$ , given by the equation

$$\frac{dx^*}{dt} = -[\nabla_{xx}^2 f(x^*(t), t)]^{-1} \nabla_{xt}^2 f(x^*(t), t), \quad x^*(t_0) = x_0^*. \quad (2)$$

By the continuity of  $\nabla_{xx}^2 f(x, t)$ , in a neighborhood of  $t_0$  the condition  $\nabla_{xx}^2 f(x^*(t), t) > 0$  is satisfied, which is a sufficient extremum condition, i.e.,

$$x^*(t) = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f(x, t). \quad (3)$$

In other words, if the minimum point for one  $t = t_0$  is known, then one can find from (2) the minimum points for the close values of  $t$ . If  $f(x, t)$  is strongly convex in  $x$  for each  $t$ , then the global minimum  $x^*(t) = \underset{x}{\operatorname{argmin}} f(x, t)$  exists and is unique for all  $t$ , and is defined by (2), which has a solution extendable to the entire axis. Thus, if the form of  $f(x, t)$

is known (it suffices that its derivatives  $\nabla_{xx}^2 f(x, t)$  and  $\nabla_{tx}^2 f(x, t)$  are admissible), then the trajectory of the minimum points  $x^*(t)$  can be tracked by solving the differential equation (2), provided the minimum point is known at any time  $t_0$ .

Of course, this approach is, chiefly, of theoretical significance, because (a) the differential equation (2) cannot be solved exactly, (b) the minimum point  $x_0^*$  at  $t_0$  can be found only approximately, and (c) the form of dependence of  $f$  on  $t$  is usually unknown. To overcome the first two drawbacks, one may go to the discrete time, i.e., replace the differential equation by a finite-difference equation and, also, choose as an initial approximation not necessarily a minimum point.

### 6.3.2 The Form of $f(x, t)$ is Unknown

Consider a somewhat different situation, without any information about the law of variation of an objective function in time. Suppose that at the  $k$ th instant of time (a discrete variant of the problem!), we have a function  $f_k(x)$ : the values of the function and of the derivatives can be computed at an arbitrary point. Then it is possible to make several iterations of some method for minimizing  $f_k(x)$  and take the resulting point as the initial approximation for minimizing  $f_{k+1}(x)$ . In the simplest variant, it is possible to make only one step of the gradient method

$$x^{k+1} = x^k - \gamma \nabla f_k(x^k), \quad (4) \quad \text{VK}$$

or, of Newton's method

$$x^{k+1} = x^k - [\nabla^2 f_k(x^k)]^{-1} \nabla f_k(x^k). \quad (5)$$

We are interested to investigate the behavior of similar iterations, or, in other words, to analyze the gradient method or Newton's method in a non-stationary case.

We examined before similar problems when we studied the influence of noise on the optimization methods. For example, if there exists a limit function  $f(x)$  such that  $f_k(x) \rightarrow f(x)$ ,  $\nabla f_k(x) \rightarrow \nabla f(x)$ , then  $\nabla f_k(x^k)$  can be written in the form  $\nabla f_k(x^k) = \nabla f(x^k) + (\nabla f_k(x^k) - \nabla f(x^k))$  and the last term can be viewed as "noise." Then (4) is but the gradient method for minimizing  $f(x)$  in the presence of noise, and the results of Section 4.2 are applicable. If there is no limit function, then methods (4), (5) have to be examined directly. To illustrate, let all the functions  $f_k(x)$  be twice differentiable and let

$$\ell I \leq \nabla^2 f_k(x) \leq L I, \quad \ell > 0, \quad (6) \quad \text{LJ}$$

for all  $x$  and  $k$ . Then each  $f_k(x)$  has a unique minimum point  $x_k^*$ . Suppose these minimum points drift with bounded rate:

$$\|x_k^* - x_{k+1}^*\| \leq a. \quad (7)$$

**THEOREM 1.** Under the assumptions made above, for method (4) with  $0 < \gamma < 2/L$  we have the bound

$$\overline{\lim_{k \rightarrow \infty}} \|x^k - x_k^*\| \leq \frac{a}{1-q}, \quad q = \max \{|1-\gamma\ell|, |1-\gamma L|\} < 1. \quad (8)$$

**PROOF.** As in proving Theorem 3 in Section 1.4, we have

$$\|x^{k+1} - x_k^*\| = \|x^k - \gamma \nabla f_k(x^k) - x_k^*\| \leq q \|x - x_k^*\|$$

yielding

$$\|x^{k+1} - x_{k+1}^*\| \leq \|x^{k+1} - x_k^*\| + \|x_{k+1}^* - x_k^*\| \leq q \|x - x_k^*\| + a.$$

Using Lemma 1 of Section 2.2 for  $u_k = \|x^k - x_k^*\|$ , we get (8).  $\square$

Thus the gradient method (4) tracks the nonstationary minimum with accuracy to within quantities of order  $a$ . Without any information about the law of the motion of a minimum, one should not expect anything more.

In some cases the information might be obtainable. For example, it can be known that the trajectory of the optima is described by the difference equation

$$x_{k+1}^* = g_k(x_k^*), \quad (9)$$

where the initial value  $x_0^*$  is unknown (cf. the description of the continuous trajectory  $x^*(t)$  using (2)). In this case, it is appropriate to introduce the prediction given by (9) into the minimization methods. In particular, the gradient method (4) takes the form

$$x^{k+1} = g_k(x^k) - \gamma \nabla f_k(g_k(x^k)). \quad (10)$$

### 6.3.3 Summary

We began our analysis of optimization methods with a simple case—a nonsingular unconstrained minimum of a smooth function with complete information on the problem—and gradually incorporated into our analysis all possible complicating factors, such as unadmissibility of derivatives,

noise, nonsmoothness of the function, singularity of the minimum, multimodality, nonstationary problems. One should not conclude, however, that we have exhausted by any means all aspects of the problem of unconstrained minimization. The variety of practical optimization problems is so enormous that they go beyond even the most general schemes. In particular, we have ignored so far methods for minimizing functions of special form. We shall describe several of these methods Part III, where we give concrete examples of optimization problems.



## **PART II**

# **CONSTRAINED MINIMIZATION**

## CHAPTER 7

### MINIMIZATION ON SIMPLE SETS

We begin the study of constrained minimization problems with the simplest ones having the form

$$\min_{x \in Q \subset \mathbb{R}^n} f(x), \quad (\text{A})$$

where  $Q$  is a set of “simple structure.” In principle, the conditions imposed on this set in the theorems given below are very general (convexity, closedness, etc.). However, these results become meaningful only if one can find in a simple way for  $Q$  the objects mentioned in the theorems (support hyperplane, projection, etc.). That is what is meant by the term *simple set*. The parallelepiped  $Q = \{x: a \leq x \leq b\}$ , the ball  $Q = \{x: \|x\| \leq \alpha\}$ , the linear manifold  $Q = \{x: Ax = b\}$ , are good examples. The constraints given by such sets are frequently shaped either by the physical nature of the variables (e.g., the requirement for nonnegativity) or by *a priori* information concerning the solution.

## 7.1 THEORETICAL FOUNDATIONS

### 7.1.1 Extremum Conditions in the Smooth Case

The point  $x^* \in Q$  is said to be a *local minimum point* (or simply, a minimum point) in problem (A) if  $f(x) \geq f(x^*)$  for all  $x \in Q$ ,  $\|x - x^*\| \leq \varepsilon$  for some  $\varepsilon > 0$ . If  $f(x) \geq f(x^*)$  for all  $x \in Q$ , it is called a *global minimum*.

**THEOREM 1** (necessary first-order minimum condition). Let  $f(x)$  be differentiable at the minimum point  $x^*$ , and let  $Q$  be a convex set. Then

$$(\nabla f(x^*), x - x^*) \geq 0 \quad \text{for all } x \in Q. \quad (1)$$

**PROOF.** Let  $(\nabla f(x^*), x^0 - x^*) < 0$  for some  $x^0 \in Q$ . Then  $x(\alpha) = x^* + \alpha(x^0 - x^*) \in Q$  for  $0 \leq \alpha \leq 1$  by the convexity of  $Q$  and

$$f(x(\alpha)) = f(x^*) + \alpha(\nabla f(x^*), x^0 - x^*) + o(\alpha) < f(x^*)$$

for sufficiently small  $\alpha > 0$ , which contradicts the local optimality of  $x^*$ .  $\square$

A vector  $a \in \mathbf{R}^n$  satisfying the condition  $(a, x - x^*) \leq 0$  for all  $x \in Q$  is said to support  $Q$  at the point  $x^* \in Q$  (if  $a \neq 0$ , then it defines a supporting hyperplane  $(a, x - x^*) = 0$ , cf. Section 5.1). Hence condition (1) can be stated differently: the vector  $-\nabla f(x^*)$  supports  $Q$  at the local minimum point  $x^*$ . Furthermore, every vector of the form  $s = x - x^*$ ,  $x \in Q$ , is said to be a feasible direction at the point  $x^*$  relative to the convex set  $Q$ . This term derives from the fact that  $x^* + \alpha s \in Q$  for all  $0 \leq \alpha \leq 1$ . Recalling formula (6) in Section 1.1 for the directional derivative  $f'(x; s) = (\nabla f(x), s)$ , we can formulate the extremum condition as the following: the derivative in any feasible direction at a minimum point is nonnegative.

The geometric meaning of (1) is very simple (Fig. 29): the sets  $Q$  and  $S = \{x: (\nabla f(x^*), x - x^*) < 0\}$  formed by the directions of local decrease of  $f(x)$  at  $x^*$  must not intersect.

In contrast to unconstrained minimization problems, for problem (A) a sufficient extremum condition can be formulated in terms of the first derivative for a nonconvex  $f(x)$ .

**THEOREM 2** (sufficient first-order minimum condition). Let  $f(x)$  be differentiable at the point  $x^* \in Q$ , let  $Q$  be convex and let the condition

$$(\nabla f(x^*), x - x^*) \geq \alpha \|x - x^*\|, \quad \alpha > 0, \quad (2)$$

be satisfied for all  $x \in Q$ ,  $\|x - x^*\| \leq \varepsilon$ ,  $\varepsilon > 0$ . Then  $x^*$  is a local minimum point of  $f(x)$  on  $Q$ .

**PROOF.** Take  $\varepsilon_1 > 0$ ,  $\varepsilon_1 \leq \varepsilon$ , so that

$$|f(x^* + y) - f(x^*) - (\nabla f(x^*), y)| \leq \alpha \|y\|/2$$

for  $\|y\| \leq \varepsilon_1$ . Then for  $x \in Q$ ,  $\|x - x^*\| \leq \varepsilon_1$ , we have

$$f(x) \geq f(x^*) + (\nabla f(x^*), x - x^*) - \alpha \|x - x^*\|/2 \geq f(x^*) + \alpha \|x - x^*\|/2,$$

i.e.,  $x^*$  is a local minimum point.  $\square$

Observe that (2) cannot hold if  $x^*$  is an interior point of  $Q$ , and hence under the conditions of Theorem 2 the minimum is necessarily attained at a boundary point of  $Q$  (Fig. 30).

In terms of the directional derivative, (2) can be written as

$$f'(x^*; s) \geq \alpha \|s\|, \quad \alpha > 0, \quad (3)$$

for all feasible  $s$ . Note that the sufficient extremum condition of the form " $f'(x^*; s) > 0$  for all feasible  $s$ " is actually false (Exercise 1).

Let us make the extremum conditions more precise for several important examples of sets  $Q$ .

Let

$$Q = \{x \in \mathbf{R}^n : a \leq x \leq b\}. \quad (4)$$

Then from Theorems 1 and 2 for an  $f(x)$  differentiable at  $x^* \in Q$  we obtain that if  $x^*$  is a minimum point of  $f(x)$  on  $Q$ , then

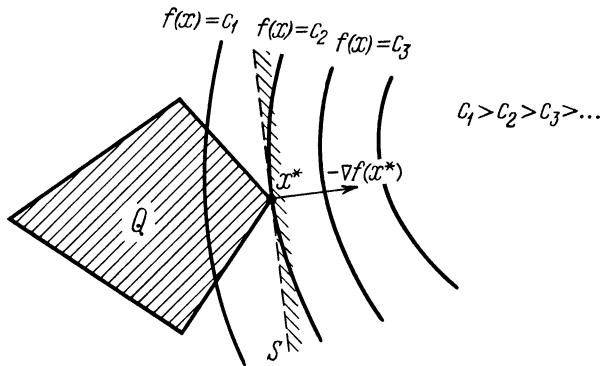


Fig. 29 Extremum conditions on the set  $Q$ .

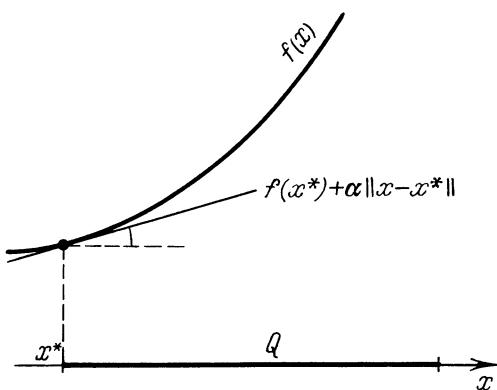


Fig. 30 The sharp minimum in a constrained problem.

$$\frac{\partial f(x^*)}{\partial x_i} \begin{cases} = 0, & a_i < x_i^* < b_i, \\ \geq 0, & x_i^* = a_i, \\ \leq 0, & x_i^* = b_i, \end{cases} \quad (5)$$

and if  $x_i^* = a_i$  or  $x_i^* = b_i$  for all  $1 \leq i \leq n$  and

$$\frac{\partial f(x^*)}{\partial x_i} \begin{cases} > 0, & x_i^* = a_i, \\ < 0, & x_i^* = b_i, \end{cases} \quad (6)$$

then  $x^*$  — minimum point of  $f(x)$  on  $Q$ . In particular, if minimum of  $f(x), x \in \mathbf{R}^1$  is to be found, under constraint  $x \geq 0$ , then condition  $f'(0) \geq 0$  is necessary, and  $f'(0) > 0$  — sufficient for minimum in 0. In principle with (5) one can find minimum points on a parallelepiped  $Q$  by direct search: split index set  $I = \{1, \dots, n\}$  into three subsets  $I = I_0 \cup I_+ \cup I_-$ , put  $x_i = a_i$  for  $i \in I_+$ ,  $x_i = b_i$  for  $i \in I_-$ , and solve a system of equations  $\partial f(x)/\partial x_i = 0, i \in I_0$ . If the solution point  $x^*$  has  $a_i < x_i^* < b_i, i \in I_0$  and  $\partial f(x^*)/\partial x^* \geq 0, i \in I_+, \partial f(x^*)/\partial x^* \leq 0, i \in I_-$ , then in  $x^*$  necessary extremum conditions hold. Of course, this approach is not a realistic method for finding solution. Later we describe much more effective minimization methods, which rely on extremum conditions.

As second example let's consider a minimization problem on a linear manifold

$$Q = \{x \in \mathbf{R}^n : Ax = b\}, \quad (7)$$

where  $b \in \mathbf{R}^m$ ,  $A$  is a matrix  $m \times n$ . From Theorem 1 it follows that  $(\nabla f(x^*), x - x^*) \geq 0$  for all  $x \in Q$ , i.e.  $(\nabla f(x^*), s) \geq 0$  for all  $s \in L = \{s : As = 0\}$ . If there exists  $s^0 \in L$  such that  $(\nabla f(x^*), s^0) > 0$ , then  $(\nabla f(x^*), -s^0) < 0$ , which is a contradiction with  $-s^0 \in L$ . Thus  $(\nabla f(x^*), s) = 0$  for all  $s \in L$ . From here follows that (Lemma 1, Section 1, Chapter 8) there exists  $y^* \in \mathbf{R}^m$  such that

$$\nabla f(x^*) = A^T y^*. \quad (8)$$

So (8) is the necessary condition for minimum  $f(x)$  on  $Q$  in form (7).

### Exercises.

1. Consider an example in  $\mathbf{R}^2$ :  $\min(y - y^2)$ ,  $y \geq 0$ ,  $y = x_2 - x_1^2$ . Check that for  $x^* = 0$  we have  $f'(x^*; s) > 0$  for all admissible  $s$ , but  $x^*$  is not a local minimum point.

### 7.1.2 Extremum Conditions in the Convex Case

We shall use the facts from the convex function theory developed in Section 5.1.

**THEOREM 3.** Let  $f(x)$  be a convex function on  $\mathbf{R}^n$ , let  $Q$  be a convex set in  $\mathbf{R}^n$ ,  $x^* \in Q$ . Then the condition: there is a subgradient  $\partial f(x^*)$  such that for all  $x \in Q$ ,

$$(\partial f(x^*), x - x^*) \geq 0, \quad (9)$$

is necessary and sufficient that  $x^*$  be a global minimum of  $f(x)$  on  $Q$ .

**PROOF.** N e c e s s i t y. Suppose that there is no such subgradient. Then the sets  $S = \partial f(x^*)$  and  $K = \{y: (y, x-x^*) \geq 0, x \in Q\}$  do not intersect. By Lemma 6 in Section 5.1,  $S$  is convex, closed and bounded. It is easy to verify that  $K$  is convex and closed. Hence the separation theorem is applicable (Theorem 1 of Section 5.1), i.e., there are  $c \in \mathbf{R}^n$ ,  $c \neq 0$  and  $\alpha > 0$  such that  $(a, c) \leq -\alpha$  for all  $a \in S$  and  $(c, y) > 0$  for all  $y \in K$ . Let  $\Gamma$  be the closure of the cone generated by the feasible directions, i.e.,  $\Gamma = \{x: x = \lim_{k \rightarrow \infty} \lambda_k(x^k - x^*), \lambda_k > 0, x^k \in Q\}$ . If  $c \notin \Gamma$ , then we apply again the separation theorem (this is possible since  $\Gamma$  is convex and closed) and find  $b$  such that  $(b, x) \geq 0$ ,  $x \in \Gamma$ , and  $(b, c) < 0$ . Then from the definition of  $K$  and  $\Gamma$  it follows that  $b \in K$  and therefore the inequality  $(b, c) < 0$  contradicts the condition  $(c, y) \geq 0$  for all  $y \in K$ . Thus,  $c \in \Gamma$ . Therefore we can find sequences  $\lambda_k > 0$  and  $x^k \in Q$  such that  $\lambda_k(x^k - x^*) \rightarrow c$ . Take  $k$  such that

$$\|\lambda_k(x^k - x^*) - c\| \leq \alpha/(2L), \quad L = \max_{a \in S} \|a\|.$$

Then by Lemma 6 of Section 5.1,

$$\begin{aligned} f'(x^*; \lambda_k(x^k - x^*)) &= \max_{a \in S} (a, \lambda_k(x^k - x^*)) \\ &= \max_{a \in S} (a, c) + \max_{a \in S} (\lambda_k(x^k - x^*) - c, a) \leq -\alpha + \frac{1}{2}\alpha = -\frac{1}{2}\alpha. \end{aligned}$$

Hence  $f'(x^*; x^k - x^*) < 0$  and hence for sufficiently small  $\gamma > 0$  one has  $f(x^* + \gamma(x^k - x^*)) < f(x^*)$ , which is impossible if  $x^*$  is a minimum point. S u f f i c i e n c y. Let  $(\partial f(x^*), x - x^*) \geq 0$  for all  $x \in Q$  and some subgradient  $\partial f(x^*)$ . Then

$$f(x) \geq f(x^*) + (\partial f(x^*), x - x^*) \geq f(x^*)$$

for any  $x \in Q$ , i.e.,  $x^*$  is a global minimum point of  $f(x)$  on  $Q$ .  $\square$

### Exercise

- $\vdash$  iff 2. Using Theorem 3, show that  $b = P_Q(a)$  if  $(b-a, x-b) \geq 0$  for all  $x \in Q$  (cf. (5) Section 5.1). Hint:  $P_Q(a)$  is a solution of the problem  $\min_{x \in Q} \|x-a\|^2$ .

### 7.1.3 Existence, Uniqueness and Stability of a Minimum

The existence theorem differs little from Theorem 1 of Section 1.3: the boundedness condition on the set  $\{x: f(x) \leq \alpha\}$  is replaced by the boundedness condition on the set  $\{x \in Q: f(x) \leq \alpha\}$ ; otherwise the proof is the same.

**THEOREM 4** (Weierstrass). Let  $f(x)$  be a continuous function on  $Q \subset \mathbf{R}^n$ , let the set  $Q$  be closed, and let the set  $\{x \in Q: f(x) \leq \alpha\}$  be bounded and non-empty for some  $\alpha$ . Then problem (A) has a solution.  $\square$

If the sufficient minimum condition (2) is satisfied, the solution is unique.

**THEOREM 5.** Under the conditions of Theorem 2,  $x^*$  is a locally unique minimum point.

The proof follows from the inequality

$$f(x) \geq f(x^*) + \alpha \|x - x^*\|/2, \quad \|x - x^*\| \leq \varepsilon, \quad (10)$$

derived in proving Theorem 2.  $\square$

The uniqueness of a solution can be guaranteed, as before, for a strictly convex  $f(x)$ . However, other conditions can be imposed on  $Q$  which lead to a unique minimum. We call the set  $Q$  strictly convex if for any  $x_1, x_2 \in Q$ ,  $x_1 \neq x_2$ ,  $0 < \lambda < 1$  the point  $\lambda x_1 + (1-\lambda)x_2$  is an interior point of  $Q$ .

**THEOREM 6.** Let  $f(x)$  be a convex function on  $\mathbf{R}^n$ , let the set  $Q$  be strictly convex and let  $\|\partial f(x)\| \geq \varepsilon > 0$  for all subgradients and all  $x \in Q$ . Then the minimum point of  $f(x)$  on  $Q$  is unique.  $\square$

The notion of stability for problem (A) can be introduced in various ways. As earlier, we call a minimization problem (globally) *stable* if every minimizing sequence converges, i.e., if it follows from  $x^k \in Q$ ,  $f(x^k) \rightarrow f^* = \inf_{x \in Q} f(x)$  that  $x^k \rightarrow x^*$ ,  $f(x^*) = f^*$ . One can define the generalized minimizing sequence:  $f(x^k) \rightarrow f^* = \inf_{x \in Q} f(x)$ ,  $\rho(x^k, Q) \rightarrow 0$ , where  $\rho(x^k, Q) = \inf_{x \in Q} \|x^k - x\|$ , and call the problem *generalized stable* if every generalized minimizing sequence converges to the minimum point.

**THEOREM 7.** If  $f(x)$  is continuous on  $\mathbf{R}^n$ ,  $Q$  is closed and the subset  $\{x \in Q: f(x) \leq \alpha\}$  is bounded and nonempty for some  $\alpha > x$ , and the global minimum point  $x^*$  is unique, then the minimization problem is stable. If  $\{x: \rho(x, Q) \leq \varepsilon, f(x) \leq \alpha\}$  is bounded for some  $\varepsilon > 0$ , then it is generalized stable.  $\square$

One can obtain quantitative estimates of stability for strongly convex functions—these estimates are perfectly analogous to the results of Lemma 2 of Section 5.2, for the unconstrained minimum (see Exercise 6). The sharp minimum case is more interesting.

We call  $x^*$  a (global) *sharp minimum point* of  $f(x)$  on  $Q$  if for all  $x \in Q$  we have

$$f(x) \geq f(x^*) + \alpha \|x - x^*\|, \quad \alpha > 0 \quad (11)$$

(cf. (9) of Section 5.2). One can give an analogous definition of a local sharp minimum and examine likewise the more general case of a nonunique minimum, but we restrict ourselves to the simplest case. For constrained problems a sharp minimum can be attained by smooth  $f(x)$ , too (see Fig. 30).

**LEMMA 1.** The following conditions are equivalent to (11) for a convex  $f(x)$  and a convex  $Q$ :

- (i)  $f'(x^*; s) \geq \alpha \|s\|$  for all feasible directions  $s$ ;
- (ii) the set  $-\partial f(x^*)$  and the set of support vectors to  $Q$  at  $x^*$  have a common interior point.  $\square$

It follows from (i) that conditions (2) and (11) are equivalent for smooth convex functions. For nonconvex functions one can show that under the conditions of Theorem 2,  $x^*$  is a local sharp minimum point (see inequality (10)).

The fundamental property of “superstability” of a sharp minimum (see Theorem 6 in Section 5.2) is also preserved for constrained problems.

**THEOREM 8.** Let  $f(x)$  be a convex function on  $\mathbf{R}^n$ , let  $Q$  be a convex closed set, let  $x^*$  be a sharp minimum point of  $f(x)$  on  $Q$ , and let  $g(x)$  be a convex function. Then we can find an  $\varepsilon_0 > 0$  such that for  $0 < \varepsilon < \varepsilon_0$  the minimum point of the function  $f(x) + \varepsilon g(x)$  on  $Q$  is unique and coincides with  $x^*$ .  $\square$

## Exercises

3. Show that a ball is a strictly convex set, but a parallelepiped and a subspace are not.
4. Prove that if  $f(x)$  is a strictly convex function on  $\mathbf{R}^n$ , then the set  $\{x: f(x) \leq \alpha\}$  is strictly convex for any  $\alpha$ .

5. Give an example of a convex (but not strictly convex) function  $f(x)$  for which the sets  $\{x: f(x) \leq \alpha\}$  are strictly convex.
6. Let  $f(x)$  be a strongly convex function on  $\mathbf{R}^n$ , and let set  $Q$  be convex and closed. Prove that a solution  $x^*$  of problem (A) exists and is unique, and  $f(x) \geq f(x^*) + (\ell/2)\|x - x^*\|^2$  for all  $x \in Q$ , where  $\ell$  is the constant of strong convexity.
7. Investigate for which  $c \in \mathbf{R}^n$  a sharp minimum obtains in the problem  $\min \|x - c\|^2$ ,  $a \leq x \leq b$ . ANSWER: If  $c_i > b_i$  or  $c_i < a_i$  for all  $1 \leq i \leq n$ .
8. Show that  $x^*$  is a sharp minimum point under condition (6).

## 7.2 BASIC METHODS

We proceed now to investigate the basic methods for solving problem (A). These methods are similar to the gradient method, as well as to Newton's method for unconstrained minimization.

### 7.2.1 The Gradient Projection Method

This method is a direct generalization of the gradient method. Since the latter leads, in general, outside the set, it is possible to add the operation of projection onto  $Q$ . We have thus arrived at the method (Fig. 31)

$$x^{k+1} = P_Q(x^k - \gamma \nabla f(x^k)), \quad (1)$$

where  $P_Q$  is the projector onto  $Q$  (see Section 5.1).

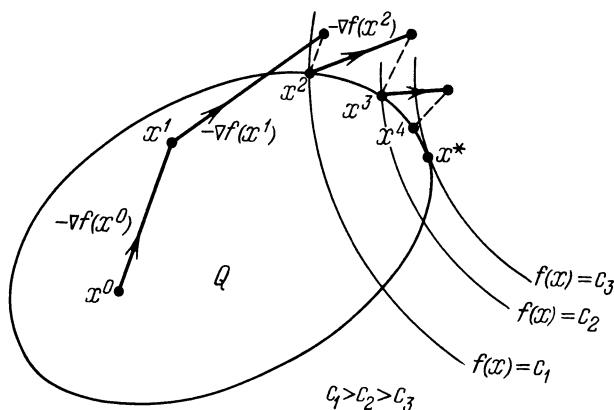


Fig. 31 The gradient-projection method.

**THEOREM 1.** Let  $f(x)$  be a convex differentiable function in  $\mathbf{R}^n$  whose gradient satisfies a Lipschitz condition with constant  $L$  on  $Q$ . Let  $Q$  be convex and closed,  $x^* = \operatorname{Argm}_{x \in Q} f(x) \neq \emptyset$  and  $0 < \gamma < 2/L$ . Then

- (i)  $x^k \rightarrow x^* \in X^*$ ;
- (ii) if  $f(x)$  is strongly convex, then  $x^k \rightarrow x^*$  with the rate of geometric progression;
- (iii) if  $f(x)$  is twice differentiable and  $\ell I \leq \nabla^2 f(x) \leq L I$ ,  $x \in Q$ ,  $\ell > 0$ , then the progression ratio is  $q = \max \{1 - \gamma\ell, 1 - \gamma L\}$ ;
- (iv) if  $x^*$  is a sharp minimum point, then the method is finite:  $x^k = x^*$  for some  $k$ .

**PROOF.** Let  $\tilde{x}$  be an arbitrary minimum point. Then taking  $x = x^k - \gamma \nabla f(x^k)$ ,  $y = \tilde{x}$  in (5) of Section 5.1, one obtains

$$(x^k - \gamma \nabla f(x^k) - x^{k+1}, \tilde{x} - x^{k+1}) \leq 0.$$

Using the minimum condition (1) of Section 7.1 yields

$$0 \geq (x^k - x^{k+1}, \tilde{x} - x^{k+1}) - \gamma (\nabla f(x^k) - \nabla f(\tilde{x}), \tilde{x} - x^{k+1}).$$

Let us transform the right-hand terms:

$$\begin{aligned} (x^k - x^{k+1}, \tilde{x} - x^{k+1}) &= (\|x^k - \tilde{x}\|^2 - \|x^k - x^{k+1}\|^2 - \|x^{k+1} - \tilde{x}\|^2)/2, \\ &= (\nabla f(x^k) - \nabla f(\tilde{x}), x^{k+1} - \tilde{x}) \\ &= (\nabla f(x^k) - \nabla f(\tilde{x}), x^k - \tilde{x}) + (\nabla f(x^k) - \nabla f(\tilde{x}), x^{k+1} - x^k) \\ &\geq L^{-1} \|\nabla f(x^k) - \nabla f(\tilde{x})\|^2 + (\nabla f(x^k) - \nabla f(\tilde{x}), x^{k+1} - x^k) \\ &\geq -(L/4) \|x^{k+1} - x^k\|^2. \end{aligned}$$

Here inequality (11) of Section 1.1 was applied first and the inequality  $\|a\|^2 + (a, b) \geq -\|b\|^2/4$  next, using  $a = L^{-1/2}(\nabla f(x^k) - \nabla f(\tilde{x}))$ ,  $b = L^{1/2} \times (x^{k+1} - x^k)$ . Hence

$$0 \geq \|x^k - \tilde{x}\|^2/2 - (1 - L\gamma/2) \|x^{k+1} - x^k\|^2/2 - \|x^{k+1} - \tilde{x}\|^2/2,$$

$$\|x^{k+1} - \tilde{x}\|^2 \leq \|x^k - \tilde{x}\|^2 - (1 - L\gamma/2) \|x^{k+1} - x^k\|^2.$$

This implies that the limit  $\lim_{k \rightarrow \infty} \|x^k - \tilde{x}\|$  exists and  $\lim_{k \rightarrow \infty} \|x^{k+1} - x^k\| = 0$ . Since the sequence  $x^k$  is bounded, it has a limit point  $x^* \in Q$ . The mapping

$T(x) = P_Q(x - \gamma \nabla f(x))$  is continuous and  $\|T(x^k) - x^k\|$  tends to 0, as was proved before. Hence  $T(x^*) = x^*$ ; but it is a sufficient minimum condition (see Exercise 1), hence  $x^* \in X^*$ . If  $\tilde{x} = x^*$ , then the entire sequence  $x^k$  converges to  $x^*$ .

To prove (ii), let us consider the mapping  $T(x)$  defined above. Applying (6) of Section 5.1, (11) of Section 1.4, and (31) of Section 1.1, we obtain

$$\begin{aligned}\|T(x) - T(y)\|^2 &\leq \|x - \gamma \nabla f(x) - y + \gamma \nabla f(y)\|^2 \\ &= \|x - y\|^2 - 2\gamma(\nabla f(x) - \nabla f(y), x - y) + \gamma^2 \|\nabla f(x) - \nabla f(y)\|^2 \\ &\leq \|x - y\|^2 - \gamma(2 - \gamma L)(\nabla f(x) - \nabla f(y), x - y) \leq q^2 \|x - y\|^2, \\ q^2 &= 1 - \gamma \ell(2 - \gamma L) < 1.\end{aligned}$$

Thus the mapping  $T(x)$  is contracting and the application of Theorem 1 of Section 2.3 yields the required result.

If  $f(x)$  is twice differentiable, we have

$$\begin{aligned}\|x^{k+1} - x^*\| &= \|T(x^k) - T(x^*)\| \leq \|x^k - x^* - \gamma(\nabla f(x^k) - \nabla f(x^*))\| \\ &= \|(I - \gamma A_k)(x^k - x^*)\| \leq q \|x^k - x^*\|, \\ A_k &= \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*)) d\tau, \quad q = \max \{|1 - \gamma \ell|, |1 - \gamma L|\}\end{aligned}$$

(cf. the proof of Theorem 3 in Section 1.4).

It remains to show that the method is finite for a sharp minimum. For arbitrary  $x \in Q$  we have, using (2) (also noting Lemma 1) of Section 7.1, that

$$\begin{aligned}(x^k - \gamma \nabla f(x^k) - x^*, x - x^*) \\ &= (x^k - x^* - \gamma(\nabla f(x^k) - \nabla f(x^*)) - \gamma(\nabla f(x^*), x - x^*) \\ &\leq (1 + \gamma L) \|x^k - x^*\| \|x - x^*\| - \gamma \alpha \|x - x^*\| \\ &= ((1 + \gamma L) \|x^k - x^*\| - \gamma \alpha) \|x - x^*\| \leq 0\end{aligned}$$

for  $\|x^k - x^*\| \leq \gamma \alpha / (1 + \gamma L)$ . Since  $x^k \rightarrow x^*$ , the last inequality holds for sufficiently large  $k$ . Applying the result of Exercise 2 of Section 7.1, we have  $x^* = P_Q(x^k - \gamma \nabla f(x^k))$ , i.e.,  $x^{k+1}$  coincides with  $x^*$ .  $\square$

We give a few examples. Let  $Q = \{x: x \geq 0\}$ ,  $x \in \mathbf{R}^n$ . Then  $P_Q(x) = x_+$  and the gradient projection method takes on the form

$$x^{k+1} = (x^k - \gamma \nabla f(x^k))_+ . \quad (2)$$

Let  $Q = \{x: a \leq x \leq b\}$ ,  $x \in \mathbf{R}^n$ . For scalars  $\tau, \alpha \leq \beta$ , we set

$$(\tau)_\alpha^\beta = \begin{cases} \tau, & \alpha \leq \tau \leq \beta, \\ \beta, & \tau > \beta, \\ \alpha, & \tau < \alpha. \end{cases} \quad (3)$$

The notation  $(x)_a^b$  has an analogous meaning for the vector  $x$ ,  $a \leq b$ ; this is the vector whose  $i$ th component is equal to  $(x_i)_a^b$ . Then the gradient projection method for the given  $Q$  has the form:

$$x^{k+1} = (x^k - \gamma \nabla f(x^k))_a^b . \quad (4)$$

Further, let  $Q$  be the ball  $Q = \{x: \|x\| \leq \rho\}$ . Then

$$x^{k+1} = \begin{cases} x^k - \gamma \nabla f(x^k) & \text{if } \|x^k - \gamma \nabla f(x^k)\| \leq \rho, \\ \rho \frac{x^k - \gamma \nabla f(x^k)}{\|x^k - \gamma \nabla f(x^k)\|} & \text{if } \|x^k - \gamma \nabla f(x^k)\| > \rho. \end{cases} \quad (5)$$

Furthermore, let  $Q$  be a linear manifold, and let  $Q = \{x \in \mathbf{R}^n: Cx = d\}$ , where  $C$  is an  $m \times n$  matrix,  $d \in \mathbf{R}^m$ . Then

$$x^{k+1} = (I - \underline{C}^+ C)(x^k - \gamma \nabla f(x^k)) + C^+ d . \quad (6) \quad \underline{C}$$

Here  $C^+$  is the pseudoinverse of  $C$  (Section 6.1). If  $x^0 \in Q$ , then

$$x^{k+1} - x^0 = T(x^k - x^0 - \gamma \nabla f(x^k)) , \quad (7)$$

where  $T = I - C^+ C$ , whereas if  $C$  is a matrix of rank  $m < n$ , then  $T = I - C^T(CCT)^{-1}C$ .

### Exercises

1. Prove that the extremum condition (1) of Section 7.1 can be written in the form:  $x^* = P_Q(x^* - \gamma \nabla f(x^*))$  for any  $\gamma > 0$ .
2. Let the sharp minimum condition in Theorem 1 be replaced by the more general condition:  $f(x) - f(x^*) \geq \alpha \|x - P_{X^*}(x)\|$ ,  $\alpha > 0$ . Prove that method (1) remains finite.

3. Show that if  $f(x)$  is not required to be convex, then  $x^k$  may not converge to a local or global minimum, but that  $f(x^{k+1}) \leq f(x^k)$  and  $\|x^{k+1} - x^k\| \rightarrow 0$ .
4. Devise a constructive rule for choosing the step size in the gradient projection method analogous to (10) of Section 3.1.
5. Give an example that demonstrates that the method  $x^{k+1} = P_Q(x^k - \gamma H \nabla f(x^k))$  does not converge for  $H \neq I$ ,  $H > 0$ .

### 7.2.2 The Subgradient Projection Method

An analogue of the subgradient method of unconstrained minimization of nonsmooth functions is the subgradient projection method

$$x^{k+1} = P_Q(x^k - \gamma_k \partial f(x^k)), \quad (8)$$

where, as before,  $\partial f(x^k)$  is any of the subgradients of the convex function  $f(x)$  at the point  $x^k$ . Rules for choosing  $\gamma_k$  are analogous to those considered in Section 5.3, and we shall mention only two very important ones:

$$\gamma_k \rightarrow 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty, \quad (9)$$

$$\gamma_k = \frac{f(x^k) - f^*}{\|\partial f(x^k)\|^2}, \quad f^* = \min_{x \in Q} f(x). \quad (10)$$

**THEOREM 2.** Let  $f(x)$  be a convex function on  $\mathbf{R}^n$ , let  $Q$  be a convex closed set, and let the set  $X^*$  of minimum points of  $f(x)$  on  $Q$  be nonempty. Then method (8), (10) converges to  $x^* \in X^*$ , and if  $\|\partial f(x)\| \leq c$  for all  $x \in Q$ , then method (8), (9) converges to  $x^* \in X^*$ , too.  $\square$

### 7.2.3 The Conditional Gradient Method

Recall that the gradient method is based on the notion of linearization. One can try to apply the same idea for the constrained problem: at the recurrent point  $x^*$  we linearize the function  $f(x)$ , then solve the problem of minimizing the linear function on  $Q$  and use the resulting point to choose the direction of motion. We then arrive at the conditional gradient method:

$$\begin{aligned} \bar{x}_k &= \operatorname{argmin}_{x \in Q} (\nabla f(x^k), x), \\ x^{k+1} &= x^k + \gamma_k (\bar{x}_k - x^k). \end{aligned} \quad (11)$$

Here it is assumed that (1) the problem of minimizing the linear function on  $Q$  has a solution (for which it is natural to require  $Q$  to be bounded),

(2) this solution can be found sufficiently simply, best of all in explicit form (see the examples in Exercise 7) and (3) it is necessary to indicate the rule for choosing  $\gamma_k$ ,  $0 \leq \gamma_k \leq 1$ .

**THEOREM 3.** let  $f(x)$  be a differentiable function whose gradient satisfies a Lipschitz condition with constant  $L$  on  $Q$ , and  $Q$  is convex, closed and bounded. Let  $\gamma_k$  be defined from the steepest descent condition:

$$\gamma_k = \underset{0 \leq \gamma \leq 1}{\operatorname{argmin}} f(x^k + \gamma(\bar{x}^k - x^k)). \quad (12)$$

Then

(i)  $(\nabla f(x^k), x^k - \bar{x}^k) \rightarrow 0$  and for every limit point of the sequence  $x^k$  the necessary extremum condition (1) of Section 7.1 is satisfied;

(ii) if  $f(x)$  is convex, then the limit points are minimum points of  $f(x)$  on  $Q$  and the following bound is true:

$$f(x^k) - f^* = O(\frac{1}{k}), \quad f^* = \underset{x \in Q}{\operatorname{min}} f(x), \quad (13) \quad \text{L 1}$$

$$f(x^k) \geq f^* \geq f(x^k) + (\nabla f(x^k), \bar{x}^k - x^k);$$

(iii) if the problem has a sharp minimum, then method (11), (12) is finite.

**PROOF.** First of all, the method is defined since under our assumptions the point  $\bar{x}^k$  exists. Let  $V(x) = f(x) - f^*$ ,  $s^k = x^k - \bar{x}^k$ . Then by the Lipschitz condition on  $\nabla V(x)$  (see (15) in Section 1.1) we have

$$V(x^{k+1}) = \underset{0 \leq \gamma \leq 1}{\operatorname{min}} V(x^k - \gamma s^k) \leq \underset{0 \leq \gamma \leq 1}{\operatorname{min}} \phi(\gamma),$$

$$\phi(\gamma) = V(x^k) - \gamma(\nabla f(x^k), s^k) + \frac{\gamma^2 L \|s^k\|^2}{2}.$$

Set  $\gamma_k^* = (\nabla f(x^k), s^k)/L \|s^k\|^2$ . By the definition of  $\bar{x}^k$ :  $(\nabla f(x^k), s^k) \geq 0$ , i.e.,  $\gamma_k^* \geq 0$ . Two cases are possible: 1)  $\gamma_k^* \leq 1$  and 2)  $\gamma_k^* > 1$ . In case 1:

$$\begin{aligned} V(x^{k+1}) &\leq \phi(\gamma_k^*) \leq V(x^k) - \frac{(\nabla f(x^k), s^k)^2}{2L \|s^k\|^2} \\ &\leq V(x^k) - \frac{(\nabla f(x^k), s^k)^2}{2LR^2}, \end{aligned} \quad (14)$$

where  $R$  is the diameter of the set  $Q$ . In case 2:  $L \|s^k\|^2 < (\nabla f(x^k), s^k)$  and

$$\begin{aligned} V(x^{k+1}) &\leq \phi(1) \leq V(x^k) - (\nabla f(x^k), s^k) + (L/2) \|s^k\|^2 \\ &\leq V(x^k) - (\nabla f(x^k), s^k)/2. \end{aligned} \quad (15)$$

Thus in both cases  $V(x^k)$  is monotonically decreasing and since  $V(x) \geq 0$ , then  $V(x^k) - V(x^{k+1}) \rightarrow 0$ . By (14) and (15) this implies that  $(\nabla f(x^k), s^k) \rightarrow 0$ .

Now let  $x^*$  be any limit point of the sequence  $x^k$  (it is known to exist since  $Q$  is bounded),  $x^{k_i} \rightarrow x^*$ . Then for any  $x \in Q$ ,

$$\begin{aligned} (\nabla f(x^*), x - x^*) &= (\nabla f(x^*) - \nabla f(x^{k_i}), x - x^*) + (\nabla f(x^{k_i}), x - x^{k_i}) \\ &\quad + (\nabla f(x^{k_i}), \bar{x}^{k_i} - x^{k_i}) + (\nabla f(x^{k_i}), x^{k_i} - x^*). \end{aligned}$$

The first and fourth terms on the right tend to zero as  $i \rightarrow \infty$  since  $x^{k_i} \rightarrow x^*$ , the second term is nonnegative by the definition of  $\bar{x}^k$ , and the third term tends to 0 by what has been proven. Hence  $(\nabla f(x^*), x - x^*) \geq 0$ , i.e., condition (1) of Section 7.1 is satisfied.

Let  $f(x)$  be convex. Then  $x^*$  is a minimum point,  $V(x^k) \rightarrow 0$  and  $V(x^k) \leq (\nabla f(x^k), x - x^*) \leq (\nabla f(x^k), x^k - \bar{x}^k)$ , i.e.,  $V(x^k) \leq (\nabla f(x^k), s^k)$ . On the other hand, from (14) and (15) we obtain

$$\begin{aligned} \checkmark \quad (\nabla f(x^k), s^k) &\leq \max \{ [2LR^2(V(x^k) - V(x^{k+1}))]^{1/2}, 2(V(x^k) - V(x^{k+1})) \} \\ \checkmark \quad &\leq [2LR^2(V(x^k) - V(x^{k+1}))]^{1/2} \end{aligned}$$

for sufficiently large  $k$  since  $V(x^k) \rightarrow 0$ . Hence  $V(x^{k+1}) \leq V(x^k) - (2LR^2)^{-1} \times V(x^k)^2$ . Using Lemma 6 of Section 2.2 yields (13).

Finally, let  $f(x)$  have a sharp minimum on  $Q$  at  $x^*$ . Then for any  $x \in Q$ , we have

$$\begin{aligned} (\nabla f(x^k), x^* - x) &= (\nabla f(x^*), x^* - x) + (\nabla f(x^k) - \nabla f(x^*), x^* - x) \\ &\leq -\alpha \|x - x^*\| + L \|x^k - x^*\| \|x - x^*\| \leq 0 \end{aligned}$$

for  $x^k$  sufficiently close to  $x^*$ . Hence for such  $x^k$  one has  $\bar{x}^k = x^*$  (by the definition of  $\bar{x}^k$ ). Since  $x^*$  is the unique minimum point, then  $\gamma_k = 1$  (see (12)) and  $x^{k+1} = x^*$ .  $\square$

Let us illustrate with an example that the bound (13) cannot be improved even for a strongly convex  $f(x)$ . Let  $x \in \mathbf{R}^2$ ,

$$f(x) = x_1^2 + (1 + x_2)^2, \quad Q = \{x: |x_1| \leq 1, 0 \leq x_2 \leq 1\} \quad (16)$$

(Fig. 32). Then  $x^* = \{0, 0\}$ ,  $\bar{x}^k = \{1, 0\}$  if  $x_1^k < 0$  and  $\bar{x}^k = \{-1, 0\}$  if  $x_1^k > 0$ . Here for all  $k$ , (14) turns into equality and we obtain  $v_{k+1} = v_k - (\frac{1}{4}) \|s^k\|^2 v_k^2$ ,  $\|s^k\|^2 \rightarrow 1$ , i.e.,  $v_k = 4/k + o(1/k)$ , where  $v_k = f(x^k) - f^* = \|x^k - x^*\|^2$ . This situation is typical: if  $Q$  is a polyhedron, while the minimum of the smooth function  $f(x)$  is attained at other points than at a vertex of  $Q$ , then the

convergence rate is just as low. This is really not surprising because only vertices of  $Q$  can be taken as  $\bar{x}^k$ . Hence the direction of motion  $\bar{x}^k - x^k$  differs strongly from the direction of motion toward the minimum  $x^* - x^k$ .

On the other hand, as was shown earlier, if the problem has a sharp minimum, the conditional gradient method is finite. Thus, the convergence rate depends on the structure of the solution as well as on the properties of  $f(x)$  and  $Q$  (smoothness, convexity, strong convexity, etc.).

The parameter  $\gamma_k$  in the conditional gradient method can also be chosen differently than as in (12) (Exercise 9). However, the simplest way,  $x^{k+1} = \bar{x}^k$  (i.e.,  $\gamma_k = 1$ ) won't do.

Furthermore, the conditional gradient method does not extend to non-smooth problems. The reason is that the minimum point of  $f(x)$  on  $Q$  is not a fixed point of a method of the form (11), in which the gradient is replaced by an arbitrary subgradient.

### Exercises

6. When is there a solution for the following problems?

- (a)  $\min (c, x), x \geq 0, x \in \mathbf{R}^n$ ;
- (b)  $\min (c, x), Ax = b, x \in \mathbf{R}^n$ ;
- (c)  $\min (c, x), (Ax, x) \leq \beta, \beta > 0, A \geq 0$ ?

ANSWER: (a) if  $c \geq 0$ ; (b) if  $A^T c = 0$ ; (c) if  $(c, e^i) = 0$  for all eigenvectors  $e^i$  of the matrix  $A$  corresponding to the null eigenvalues.

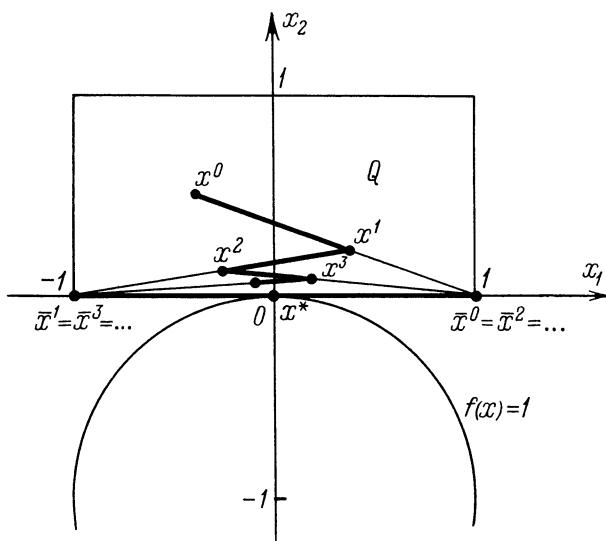


Fig. 32 Slow convergence of the ~~constrained~~ gradient method.

→ conditional

7. Verify that the solution of the following elementary minimization problem is correct:

$$(a) \quad x^* = \underset{a \leq x \leq b}{\operatorname{argmin}} (c, x), \quad x_i^* = \begin{cases} a_i & \text{if } c_i > 0, \\ b_i & \text{if } c_i < 0, \\ \text{anything between } a_i \text{ and } b_i & \text{if } c_i = 0; \end{cases}$$

$$(b) \quad x^* = \underset{\|x\| \leq \rho}{\operatorname{argmin}} (c, x) = -\frac{c\rho}{\|c\|};$$

$$(c) \quad x^* = \underset{(1/2)(Ax, x) - (bx) \leq \alpha}{\operatorname{argmin}} (c, x), \quad A > 0, \alpha > 0, \quad x^* = A^{-1}(b - \lambda c),$$

where  $\lambda$  is found from the equation  $(Ax^*, x^*)/2 - (b, x^*) = \alpha$ .

8. Let  $x^*$  be the solution of the problem  $\min (c, x)$ ,  $x \in Q$ . Prove that  $x^* = \lim P_Q(\lambda c)$  as  $\lambda \rightarrow \infty$ .

9. Prove that all the assertions of Theorem 3 remain valid if the step size is chosen from the condition

$$\gamma_k = \min \{1, (\nabla f(x^k), x^k - \bar{x}^k)/L \|x^k - \bar{x}^k\|^2\}.$$

10. We call the set  $Q$  strongly convex if there is a  $\beta > 0$  such that if  $x, y \in Q$ , then  $z \in Q$  for

$$\|z - (x+y)/2\| \leq \beta \|x - y\|^2.$$

Show that if  $f(x)$  is a strongly convex function on  $\mathbf{R}^n$ , then the set  $Q_\alpha = \{x: f(x) \leq \alpha\}$  is strongly convex. Prove that a strongly convex set other than  $\mathbf{R}^n$  is bounded.

11. Prove that if  $f(x)$  is convex and  $\|\nabla f(x)\| \geq \varepsilon > 0$  for  $x \in Q$ , while  $Q$  is strongly convex, then the conditional gradient method under the conditions of Theorem 3 converges linearly.

12. Introduce the function

$$\psi(x) = \underset{y \in Q}{\operatorname{min}} [f(x) + (\nabla f(x), y - x)].$$

Show that if  $f(x)$  is convex, then  $\psi(x) \leq f(x)$  for all  $x \in Q$ , and equality obtains iff  $x = \underset{x' \in Q}{\operatorname{argmin}} f(x')$ . Try to investigate the properties of the function  $\psi(x)$  (convexity, differentiability, etc.) Think about how the conditional gradient method can be interpreted in terms of the function  $\psi(x)$ .

#### 7.2.4 Newton's Method

To construct Newton's method in problem (A), one can use the same idea of quadratic approximation of  $f(x)$  as for the case of unconstrained mini-

um. The only difference is that it is necessary to find the approximation minimum on the set  $Q$  rather than over the entire space. These arguments lead to the method

$$\begin{aligned} x^{k+1} &= \underset{x \in Q}{\operatorname{argmin}} f_k(x), \\ f_k(x) &= f(x^k) + (\nabla f(x^k), x - x^k) + (\nabla^2 f(x^k)(x - x^k), x - x^k)/2. \end{aligned} \quad (17)$$

**THEOREM 4.** Let  $f(x)$  attain a minimum on a closed convex set  $Q$  at a point  $x^*$ , at which  $f(x)$  is twice differentiable on  $Q$  in a neighborhood of  $x^*$ , let  $\nabla^2 f(x)$  satisfy a Lipschitz condition, and let

$$\nabla^2 f(x^*) > 0. \quad (18)$$

Then method (17) converges locally to  $x^*$  with quadratic rate.

**PROOF.** At  $x^{k+1}$  the necessary minimum condition for  $f_k(x)$  is satisfied on  $Q$ , i.e.,

$$(\nabla f_k(x^{k+1}), x - x^{k+1}) = (\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k), x - x^{k+1}) \geq 0$$

for all  $x \in Q$  and in particular for  $x = x^*$ . Hence

$$\begin{aligned} 0 &\leq (\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1} - x^k), x^* - x^{k+1}) \\ &= (\nabla f(x^*), x^* - x^{k+1}) \\ &\quad + (\nabla f(x^k) - \nabla f(x^*), x^* - x^{k+1}). \end{aligned}$$

The first term is nonpositive by virtue of (1) of Section 7.1. Then for  $\nabla f(x^k) - \nabla f(x^*)$  we have the bound

$$\nabla f(x^k) - \nabla f(x^*) = \nabla^2 f(x^k)(x^k - x^*) + r, \quad \|r\| \leq (L/2) \|x^k - x^*\|^2$$

(see (15) of Section 1.1). Hence

$$\begin{aligned} 0 &\leq (\nabla^2 f(x^k)(x^k - x^*) + r, x^* - x^{k+1}) \\ &\leq -\ell \|x^{k+1} - x^*\|^2 + (L/2) \|x^k - x^*\|^2 \|x^{k+1} - x^*\|. \end{aligned}$$

Hence we have used the fact that  $\nabla^2 f(x^k) \geq \ell I$ ,  $\ell > 0$ , for all  $x^k$  sufficiently close to  $x^*$ , by (18) and a Lipschitz condition on the Hessian.

Therefore, either  $x^{k+1} = x^*$  or

$$\|x^{k+1} - x^*\| \leq L \|x^k - x^*\|^2 / (2\ell). \quad (19)$$

 If  $L \|x^k - x^*\| / (2\ell) < 1$ , then it follows from (19) that all the  $x^k$  remain in the same neighborhood of  $x^*$ , while the bound (19) implies the quadratic rate of convergence.  $\square$

For the case of a sharp minimum, it is not hard to prove that the method is finite. However, then it hardly makes sense to apply Newton's method, since other much simpler methods (the gradient method and the conditional gradient method) are also finite.

Newton's method can be applied only when the problem of minimizing a quadratic function on  $Q$  is easily solved. If  $Q$  is a polyhedron, then (17) is a general problem of quadratic programming. As we shall show in Chapter 10, finite algorithms are available in order to solve this problem. For the special case when  $Q$  is a parallelepiped, problem (17) can be solved using a modification of the conjugate gradient method, described in the next section. In the simplest case, when  $Q$  is a ball or a linear manifold, (17) has a quite simple solution.

## Exercises

13. Show that the solution of the problem

$$\min [(Ax, x)/2 - (b, x)], \quad A > 0, \quad \|x\| \leq \rho,$$

is the point  $(A + \lambda I)^{-1} b$ , at which  $\lambda = 0$  if  $\|A^{-1} b\| \leq \rho$ , and otherwise if  $\lambda$  is found from the equation  $\|(A + \lambda I)^{-1} b\| = \rho$ .

14. Consider modifications of Newton's method analogous to those in Section 3.1, for unconstrained minimization. Show that they converge globally.

## 7.3 OTHER METHODS

### 7.3.1 Quasi-Newton Methods

Note that all of the methods for solving smooth problems described in Section 7.2.4 can be derived by a general scheme. Let

$$x^{k+1} = \underset{x \in Q}{\operatorname{argm i n}} ((\nabla f(x^k), x - x^k) + \frac{1}{2} (H_k(x - x^k), x - x^k)), \quad (1)$$

where  $H_k \geq 0$  is a matrix.

Obviously, for  $H_k = \nabla^2 f(x^k)$ , method (1) turns into Newton's method, whereas for  $H_k = \gamma^{-1}I$  it becomes the gradient projection method since the latter can be written in the form

$$x^{k+1} = \underset{x \in Q}{\operatorname{argmin}} \|x - (x^k - \gamma \nabla f(x^k))\|^2.$$

It is possible to extend the class of methods (1), introducing the one-dimensional procedure:

$$\begin{aligned}\bar{x}^k &= \underset{x \in Q}{\operatorname{argmin}} ((\nabla f(x^k), x - x^k) + \frac{1}{2} (H_k(x - x^k), x - x^k)) , \\ x^{k+1} &= x^k + \gamma_k s^k , \quad s^k = \bar{x}^k - x^k , \quad \gamma_k = \underset{0 \leq \gamma \leq 1}{\operatorname{argmin}} f(x^k + \gamma s^k) .\end{aligned}\tag{2}$$

In particular, for  $H_k = 0$ , from (2) we obtain the conditional gradient method. Such methods require a special analysis for convergence. For example, the results on convergence of unconstrained minimization methods like  $x^{k+1} = x^k - \gamma_k H_k \nabla f(x^k)$  with arbitrary  $H_k > 0$  (Lemma 1 in Section 3.3) cannot be used as it was done in proving Theorem 1 in Section 7.2. In Lemma 1 of Section 3.3, the Lyapunov function is  $f(x) - f(x^*)$  rather than the distance to the minimum; hence it is not possible to claim that the projection operator is a relaxation (Exercise 1). It is however possible to prove the method using the same scheme as that for Theorem 4 in Section 7.2. Here is a typical result.

**THEOREM 1.** Let  $f(x)$  be twice differentiable and let  $\ell I \leq \nabla^2 f(x) \leq L I$ ,  $\ell > 0$  for all  $x \in Q$ ,  $Q$  being closed and convex, and

$$\|H_k - \nabla^2 f(x^k)\| \leq \varepsilon < \ell/2 .\tag{3}$$

Then in method (1) the  $x^k$  converges locally to  $x^* = \underset{x \in Q}{\operatorname{argmin}} f(x)$  with the rate of geometric progression, whereas if

$$\|H_k - \nabla^2 f(x^k)\| \rightarrow 0 ,\tag{4}$$

the rate is superlinear.

**PROOF.** From the definition of  $x^{k+1}$  we have

$$\begin{aligned}0 &\leq (\nabla f(x^k) + H_k(x^{k+1} - x^k), x^* - x^{k+1}) \\ &\leq (\nabla f(x^k) - \nabla f(x^*) + H_k(x^{k+1} - x^k), x^* - x^{k+1}) .\end{aligned}$$

But

$$\nabla f(x^k) - \nabla f(x^*) = \nabla^2 f(x^k)(x^k - x^*) + r, \quad \|r\| \leq \frac{1}{2} L \|x^k - x^*\|^2,$$

and hence

$$\begin{aligned} 0 &\leq ((\nabla^2 f(x^k) - H_k)(x^k - x^*) + H_k(x^{k+1} - x^*) + r, x^* - x^{k+1}) \\ &\leq \varepsilon \|x^k - x^*\| \|x^{k+1} - x^*\| - (\ell - \varepsilon) \|x^{k+1} - x^*\|^2 \\ &\quad + (L/2) \|x^k - x^*\|^2 \|x^{k+1} - x^*\|, \\ \|x^{k+1} - x^*\| &\leq \frac{\varepsilon + (L/2) \|x^k - x^*\|}{\ell - \varepsilon} \|x^k - x^*\|. \end{aligned}$$

If  $(L/2) \|x^0 - x^*\| < \ell - 2\varepsilon$ , then  $x^k \rightarrow x^*$  with the rate of geometric progression: the smaller  $\varepsilon$  the smaller the ratio. Similarly, by (4) we have

$$\|x^{k+1} - x^*\| \leq q_k \|x^k - x^*\|, \quad q_k \rightarrow 0 \quad \text{as } k \rightarrow \infty,$$

which implies superlinear convergence.  $\square$

Theorem 1 shows that in problems where the calculation of  $\nabla^2 f(x^k)$  is impossible or too laborious, an approximation of the Hessian is advantageous. This can be done just as in quasi-Newton unconstrained minimization methods, employing the data obtained in the preceding iterations. To be precise, if the gradients at the preceding points are available, then  $H$  can be reconstructed from the approximate equalities

$$\nabla f(x^{i+1}) - \nabla f(x^i) \approx H(x^{i+1} - x^i), \quad i = k, \dots, k-n+1, \quad (5)$$

provided the  $x^i$  do not lie on the same subspace.

We do not elaborate on these methods because they chiefly use the same technique as in unconstrained minimization. The only difference here is that in the constrained problem the vector  $\nabla f(x^{k+1})$  is not generally orthogonal to the direction of motion  $x^{k+1} - x^k$ .

### Exercises

1. Give an example of a function  $f(x)$  and a matrix  $H > 0$ , where the method  $x^{k+1} = x^k - H \nabla f(x^k)$  converges but  $\|x^k - x^*\|$  does not decrease monotonically ( $x^*$  is the minimum point of  $f(x)$ ). Use this example to construct an example of divergence of method (1) with  $H_k \equiv H$ .

2. Prove global convergence for method (1) when  $H_k$  is sufficiently close to  $\gamma^{-1}I$ ,  $0 < \gamma < 2/L$ .

3. Show that the method

$$x^{k+1} = P_Q(x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)) \quad \checkmark$$

does not generally converge. In particular, if  $f(x)$  is quadratic, then for any  $x^0$  the method stops at  $x^L$  (identical for all  $x^0$ ), and  $x^L$  is not generally a solution.

L 12  
v(0)

### 7.3.2 The Conjugate Gradient Method

We consider first the case when  $f(x)$  is a quadratic function, and  $Q$  is a subspace in  $\mathbb{R}^n$ ,  $Q = \{x: Cx = 0\}$ ,  $C$  is an  $m \times n$  matrix of rank  $m$ . The projection of a vector onto this subspace is given by

$$P_Q(x) = (I - C^T C)x = (I - C^T (CC^T)^{-1} C)x.$$

Let us write the conjugate gradient method in which the vector  $\nabla f(x)$  is replaced by its projection onto  $Q$ :

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k p^k, \quad \alpha_k = \underset{\alpha}{\operatorname{argmin}} f(x^k + \alpha p^k), \quad x^0 \in Q, \\ p^k &= -P_Q \nabla f(x^k) + \beta_k p^{k-1}, \quad p_0 = -P_Q \nabla f(x^0), \\ \beta_k &= \|P_Q \nabla f(x^k)\|^2 / \|P_Q \nabla f(x^{k-1})\|^2. \end{aligned} \quad (6)$$

One can show (Exercise 4) that if  $f(x)$  is a quadratic function,  $f(x) = (Ax, x)/2 - (b, x)$ , and  $(Ax, x) \geq \alpha \|x\|^2$ ,  $\alpha > 0$ , for all  $x \in Q$ , then method (6) stops in no more than  $n-m$  steps.

Thus the conjugate gradient method remains finite when a quadratic function is minimized on a subspace, the number of steps is smaller the more constraints there are. Of course, each iteration of the method involves the additional calculations of the projection onto a subspace.

Now let  $Q$  be the positive orthant in  $\mathbb{R}^n$ , i.e.,  $Q = \{x: x \geq 0\}$ , and let  $f(x)$  be quadratic as before. Then its minimization on  $Q$  can be reduced to sequential minimization on the faces of the  $Q$ . These faces have the form  $\{x_i = 0, i \in I, x_i > 0, i \notin I\}$ , where  $I$  is some set of indices in  $\{1, \dots, n\}$ . Minimization on the subspace  $L = \{x: x_i = 0, i \in I\}$  is simple—it is necessary to make calculations as in the conjugate gradient method, changing to zeros the components from the set  $I$  both for the vectors  $x^k$  and the gradients  $\nabla f(x^k)$  (see (6) and Exercise 5). Taking these considerations into account, we arrive at a method for minimizing  $f(x)$  on  $Q$ , which in the coordinate form is:

$$\checkmark(7) \quad x^{k+1} = x^k + \alpha_k p^k, \quad \alpha_k = \underset{\substack{\alpha \geq 0 \\ x^k + \alpha p^k \geq 0}}{\operatorname{argmin}} f(x^k + \alpha p^k), \quad \checkmark$$

$$p_i^k = \begin{cases} -\nabla f(x^k)_i + \beta_k p_i^{k-1}, & i \in I_k, \\ 0, & i \notin I_k, \end{cases}$$

$$\beta_k = \begin{cases} \sum_{i \in I_k} (\nabla f(x^k)_i)^2 / \sum_{i \in I_k} (\nabla f(x^{k-1})_i)^2 & \text{if } I_k = I_{k-1}, \\ 0 & \text{if } k = 0 \text{ or } I_k \neq I_{k-1}, \end{cases}$$

$$I_k = \begin{cases} \{i: x_i^k = 0, \nabla f(x^k)_i > 0\} & \text{if } k = 0 \text{ or } \nabla f(x^k)_i = 0 \text{ for all } i \in I_{k-1}, \\ I_{k-1} \cup \{i: x_i^k = 0\} & \text{otherwise.} \end{cases}$$

In other words,  $f(x)$  is minimized by the conjugate gradient method on the set  $L_k = \{x: x_i = 0, i \in I_k, x_i > 0, i \notin I_k\}$ . The process stops either when one of the components (not belonging to  $I_k$ ) of  $x^k$  vanishes (in this case the index of this component is added to the set  $I_k$ ), or when the minimum on  $L_k$  is found (in this case the set  $I_k$  is innovated). It is possible to show that if  $f(x) = (Ax, x)/2 - (b, x)$ ,  $A > 0$ , then this method is finite.

We have obtained the finite method for solving the problem of minimizing a quadratic function under the constraints  $x \geq 0$ . Of course other finite variants of the conjugate gradient method are possible. Also, this method can be extended to the case involving constraints of the form  $a \leq x \leq b$ .

The same idea can be used in order to minimize a nonquadratic function on an orthant or on a parallelepiped. In that case, one needs to regulate the accuracy of a solution of the minimization problem on a face. In general, such methods are not finite.

### Exercises

4. Let  $Q$  be a subspace in  $\mathbf{R}^n$ , let  $f(x)$  be a differentiable function on  $\mathbf{R}^n$ . Let  $f_Q(x)$  denote its restriction to  $Q$ . Then the gradient  $\nabla f_Q(x)$  at  $x \in Q$  is defined by the equality  $\nabla f_Q(x+y) = f_Q(x) + (\nabla f_Q(x), y) + o(y)$  for all  $y \in Q$ , where  $\nabla f_Q(x) \in Q$ . Prove that  $\nabla f_Q(x) = P_Q \nabla f(x)$ . Using this result, show that (6) is the conjugate gradient method for unconstrained minimization of  $f_Q(x)$ . It follows that if  $f_Q(x)$  is quadratic, the method is finite.

5. Show that if  $Q = \{x: x_i = 0, i \in I\}$ , then  $P_Q(x)_i = 0$  if  $i \in I$ , and  $P_Q(x)_i$  if  $i \notin I$ .

### 7.3.3 Minimization of Nonsmooth Functions

In describing the methods of unconstrained minimization of convex nonsmooth functions in Section 5.4, we assumed that the region of localization of the minimum is specified. If this region is taken to be  $Q$ , then it turns out that all these methods are also usable for constrained problems. Thus the cutting-plane method, the Chebyshev centers method, the center-of-gravity method, and others, apply verbatim to problems (A). In this case, at each step one is solving the problem of minimizing a linear or quadratic function on a set  $Q_k$ , which is given by the condition  $x \in Q$  as well as some additional linear constraints. If  $Q$  is a polyhedron, we obtain a problem of linear or quadratic programming, which can be solved via standard methods. All of the results related to the convergence and rate of convergence in Section 5.4 hold for constrained problems, too. We note that the presence of a sharp minimum in the nonsmooth case does not lead, in general, to the finiteness of the methods.

## 7.4 THE INFLUENCE OF NOISE

We are not going to discuss all possible cases in the same detail as we did in the unconstrained minimization problems (Chapter 4). We are mainly interested in different, new effects produced by the constraints.

### 7.4.1 Absolute Deterministic Noise

Suppose that instead of the gradient  $\nabla f(x^k)$  (or the subgradient  $\partial f(x^k)$ ) we know only their approximations  $\tilde{\nabla}f(x^k)$  ( $\tilde{\partial}f(x^k)$ ), and

$$\|\tilde{\nabla}f(x^k) - \nabla f(x^k)\| \leq \varepsilon \quad (\|\tilde{\partial}f(x^k) - \partial f(x^k)\| \leq \varepsilon). \quad (1)$$

Suppose we apply the methods of Section 7.2 in this situation, i.e., in these methods we replace  $\nabla f(x^k)$  and  $\partial f(x^k)$  by  $\tilde{\nabla}f(x^k)$  and  $\tilde{\partial}f(x^k)$ . In this case, generally, the gradient projection method and the subgradient projection method cease to converge, and lead to a neighborhood of the minimum, the size of which depends on  $\varepsilon$ . The situation is slightly different with the conditional gradient method. First of all it includes the one-dimensional minimization operation, which cannot be executed exactly. Moreover, the point  $\bar{x}^k$  can change drastically when  $\nabla f(x^k)$  is replaced by  $\tilde{\nabla}f(x^k)$ . Hence the conditional gradient method is hardly appropriate for problems involving noise.

The case of a sharp minimum presents a new situation.

**THEOREM 1.** Let  $x^*$  be a sharp minimum point of a differentiable convex function  $f(x)$  on the convex set  $Q$ . Suppose that a projection onto  $Q$  can be executed exactly. Under the conditions of Theorem 1 of Section 7.2, the gradient projection method remains finite if  $\nabla f(x^k)$  is replaced in it by  $\tilde{\nabla} f(x^k)$  and  $\varepsilon > 0$  is sufficiently small.

The proof is similar to that of Theorem 1 of Section 7.2.  $\square$

To conclude, for problems with a smooth  $f(x)$  and a sharp minimum, some methods are superstable and give an exact solution even in the presence of absolute (but sufficiently small) noise.

#### 7.4.2 Absolute Random Noise

Let the noise

$$\xi^k = \nabla f(x^k) - \tilde{\nabla} f(x^k) \quad (\xi^k = \partial f(x^k) - \tilde{\partial} f(x^k)) \quad (2)$$

be random, independent, centered, and have bounded variance:

$$E\xi^k = 0, \quad E\|\xi^k\|^2 \leq \sigma^2. \quad (3)$$

**THEOREM 2.** Let  $f(x)$  be a convex function on  $\mathbf{R}^n$ , and let  $Q$  be a bounded closed convex set. Then in the method

$$x^{k+1} = P_Q(x^k - \gamma_k \tilde{\partial} f(x^k)), \quad \sum_{k=0}^{\infty} \gamma_k^2 < \infty, \quad \sum_{k=0}^{\infty} \gamma_k = \infty, \quad (4)$$

when (2) and (3) are satisfied we have  $x^k \rightarrow x^*$  a.s., where  $x^*$  is some minimum point of  $f(x)$  on  $Q$ . If  $f(x)$  is strongly convex, then  $E\|x^k - x^*\|^2 \rightarrow 0$  (the condition  $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$  here can be changed to  $\gamma_k \rightarrow 0$ ), whereas if  $\gamma_k = \gamma/k$  and  $\gamma$  is sufficiently great, then  $E\|x^k - x^*\|^2 = O(1/k)$ .  $\square$

For sharp minimum problems there is apparently no need to make  $\gamma_k$  tend to zero. One may assume that for a correct adjustment of the step size the gradient method in the presence of noise will be finite almost surely for a sharp minimum problem.

For the conditional gradient method, at first glance it seems natural to proceed just as in method (4), i.e., replace the exact value of the gradient by an approximate value and let the step size tend to zero:

$$\begin{aligned} \bar{x}^k &= \underset{x \in Q}{\operatorname{argmin}} (\tilde{\nabla} f(x^k), x), \\ x^{k+1} &= x^k + \gamma_k (\bar{x}^k - x^k), \quad \gamma_k \rightarrow 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty. \end{aligned} \quad (5)$$

However, such a method does not generally converge. For example, assume we are seeking the minimum of a smooth function  $f(x)$ ,  $x \in \mathbb{R}^1$ , on  $Q = [-\alpha, \beta]$ ,  $\alpha > 0$ ,  $\beta > 0$ , while the minimum obtains at  $x^* = 0 \in Q$ . Then for  $x^k = x^*$  one has  $\nabla f(x^k) = 0$  and  $\bar{x}^k = -\alpha$  if  $\xi^k > 0$  and  $\bar{x}^k = \beta$  if  $\xi^k < 0$ . For symmetrically distributed noise  $E(\bar{x}^k - x^k) = (\beta - \alpha)/2 \neq 0$  for  $\beta \neq \alpha$ . Thus, at a minimum point of  $f(x)$  the mean value of the direction of motion is nonzero, and hence the method cannot converge to this point.

Convergence can be achieved in the conditional gradient method by introducing a gradient averaging procedure:

$$\begin{aligned}\bar{x}^k &= \underset{z \in Q}{\operatorname{argmin}} (y^k, z), \\ y^k &= y^{k-1} + \mu_k (\tilde{\nabla} f(x^k) - y^{k-1}), \quad \mu_k \rightarrow 0, \quad \sum_{k=0}^{\infty} \mu_k = \infty, \\ x^{k+1} &= x^k + \gamma_k (\bar{x}^k - x^k), \quad \gamma_k \rightarrow 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty.\end{aligned}\tag{6}$$

Here  $y^k$  is the value of the gradient averaged over the preceding iterations.

### 7.4.3 Relative Noise

Suppose the noise satisfies the condition

$$\|\nabla f(x) - \tilde{\nabla} f(x)\| \leq \alpha \|\nabla f(x)\|. \tag{7}$$

We saw (Theorem 2 in Section 4.2) that the gradient method is stable under such noise if the noise level is below 100% (i.e.,  $\alpha < 1$ ). In constrained problems, this is not the case: since, in general,  $\nabla f(x^*) \neq 0$  at the minimum point  $x^*$ , then the quantity  $\|\tilde{\nabla} f(x) - \nabla f(x)\|$  does not need to tend to zero when  $x$  gets close to  $x^*$ . Hence the situations with absolute and with relative noise barely differ in this case, and hence, for example, it is impossible to guarantee that the gradient projection method converges under deterministic relative noise of any level.

The real analog of relative errors for constrained problems is given by conditions such that

$$\|\nabla f(x^k) - \tilde{\nabla} f(x^k)\| \leq \alpha \|x^k - x^*\|, \tag{8}$$

$$\|\nabla f(x^k) - \tilde{\nabla} f(x^k)\| \leq \alpha \|x^{k+1} - x^*\|. \tag{9}$$

However, these conditions are somewhat artificial, and we ignore them.

## CHAPTER 8

### PROBLEMS WITH EQUALITY CONSTRAINTS

In this chapter we consider the problem

$$\begin{aligned} \min f(x), \quad x \in \mathbf{R}^n, \\ g_i(x) = 0, \quad i = 1, \dots, m, \end{aligned} \tag{A}$$

where  $f$  and  $g_i$  are smooth functions. This is a special case of a general problem of mathematical programming (see Chapter 9). We shall consider this problem in some detail since the ideas are most transparent.

## 8.1 THEORETICAL FOUNDATIONS

### 8.1.1 Lagrange Multipliers

Let  $Q = \{x: g_i(x) = 0, i = 1, \dots, m\}$ . The points  $x \in Q$  are said to be *admissible*. The point  $x^*$  is called a (local) *minimum* for problem (A) if it is admissible and  $f(x^*) \leq f(x)$  for all admissible  $x$  sufficiently close to  $x^*$ .

**THEOREM 1** (necessary first-order minimum condition). Let  $x^*$  denote a minimum point in problem (A), and let the functions  $f(x)$ ,  $g_i(x)$  be continuously differentiable in a neighborhood of  $x^*$ . Then we can find  $y_0^*$ ,  $y_1^*$ , ...,  $y_m^*$ , not all of them being equal to zero, such that

$$y_0^* \nabla f(x^*) + \sum_{i=1}^m y_i^* \nabla g_i(x^*) = 0. \tag{1}$$

We say that  $x^*$  is a *regular minimum point* if  $f(x)$ ,  $g_i(x)$  are continuously differentiable in a neighborhood of  $x^*$  and  $\nabla g_i(x^*)$ ,  $i = 1, \dots, m$ , are linearly independent.

**THEOREM 2** (The rule of Lagrange multipliers). If  $x^*$  is a regular minimum point, then we can find  $y_1^*, \dots, y_m^*$  such that

$$\nabla f(x^*) + \sum_{i=1}^m y_i^* \nabla g_i(x^*) = 0. \quad (2)$$

The  $y_1^*, \dots, y_m^*$  in (2) are called *Lagrange multipliers*. The fact that the Lagrange multipliers rule holds, in general, only under the regularity condition can easily be seen from simple examples. Thus, in the problem  $\min x$ ,  $x_2 = 0$ ,  $x \in \mathbf{R}^1$  the point  $x^* = 0$  is a minimum (but not a regular) point, and equality (2) is unsatisfiable for any  $y^*$  since  $f'(x^*) = 1$ ,  $g'(x^*) = 0$  (see also Exercise 1).

Theorem 2 follows immediately from Theorem 1. Indeed, in the regular case,  $y_0^* \neq 0$  (otherwise  $\sum_{i=1}^m y_i^* \nabla g_i(x^*) = 0$ , not all of the  $y_1^*, \dots, y_m^*$  being equal to zero, which contradicts the linear independence of the  $\nabla g_i(x^*)$ ). Dividing (1) by  $y_0^*$ , we obtain (to within the notation) relation (2). Conversely, if Theorem 2 is proven, Theorem 1 holds as well. Indeed, if  $\nabla g_i(x^*)$  are linearly independent:  $\sum_{i=1}^m \mu_i \nabla g_i(x^*) = 0$ ,  $\sum_{i=1}^m \mu_i^2 \neq 0$ , then equality (1) holds for  $y_0^* = 0$ ,  $y_i^* = \mu_i$ ,  $i = 1, \dots, m$ . Therefore it suffices to prove Theorem 2. In what follows we shall analyze three different proofs because (1) the result per se is important and (2) the ideas of these proofs are used in constructing minimization methods.

Let us compose the Lagrange function

$$L(x, y) = f(x) + (y, g(x)) = f(x) + \sum_{i=1}^m y_i g_i(x) \quad (3)$$

defined on  $\mathbf{R}^n \times \mathbf{R}^m$ . Here and below we use the vector notation  $y = (y_1, \dots, y_m)$ ,  $g(x) = (g_1(x), \dots, g_m(x))$ . Then the rule of Lagrange multipliers is:

$$L'_x(x^*, y^*) = 0, \quad L'_y(x^*, y^*) = 0, \quad (4)$$

where  $L'_x$ ,  $L'_y$  denote the derivatives with respect to the corresponding variables. The notation in form (4) is convenient in its symmetry in the variables  $x$  and  $y$ , called respectively the *primal* and *dual* variables.

1. *Proof Based on Lyusternik's Theorem.* Let  $Q$  be some subset of  $\mathbf{R}^n$ ,  $x \in Q$ . The vector  $s \in \mathbf{R}^n$  is said to be tangent to  $Q$  at the point  $x$  if for all sufficiently small  $\tau > 0$  we can find points  $x(\tau) \in Q$  such that  $\|x(\tau) - (x + \tau s)\| = o(\tau)$  (Fig. 33). If  $Q$  is convex, then every feasible

15

✓ direction (Section 7.1) is a tangent direction also, but not conversely (see Exercise 3). Obviously, the tangent vectors form a cone  $S_Q(x)$  (i.e., if  $s \in S$ , then  $\lambda s \in S$  for  $\lambda \geq 0$ ). Notice that if  $x$  is a boundary point of a ball, then the cone  $S$  is a half-space, rather than a hyperplane. Hence the term "tangent vector" has in this case a different meaning than in Geometry.

**THEOREM 3** (Lyusternik). Let  $Q = \{x \in \mathbf{R}^n : g_i(x) = 0, i = 1, \dots, m\}$ , where the  $g_i(x)$  are continuously differentiable in a neighborhood of  $x^* \in Q$ , and  $\nabla g_i(x^*)$ ,  $i = 1, \dots, m$ , are linearly independent. Then

$$S_Q(x^*) = \{s \in \mathbf{R}^n : (s, \nabla g_i(x^*)) = 0, i = 1, \dots, m\}, \quad (5)$$

i.e., the tangent vectors to  $Q$  at  $x^*$  form a subspace orthogonal to the vectors  $\nabla g_1(x^*)$ , ...,  $\nabla g_m(x^*)$ .  $\square$

To prove the Lagrange multipliers rule, we need the following lemma.

**LEMMA 1.** Let  $A$  be an  $m \times n$  matrix, and let  $L = \{x \in \mathbf{R}^n : Ax = 0\}$  and  $(c, x) \geq 0$  for all  $x \in L$ . Then  $c = A^T y$ ,  $y \in \mathbf{R}^m$  and  $(c, x) = 0$  for  $x \in L$ .

PROOF. The set  $L_1 = \{x \in \mathbf{R}^n : x \notin A^T y, y \in \mathbf{R}^m\}$  is convex and closed, being a subspace in  $\mathbf{R}^n$ . If  $c \notin L_1$ , then by the separation theorem, the point  $c$  can be strictly separated from  $L_1$ , i.e., we can find an  $a \in \mathbf{R}^n$  such that  $(a, c) < 0$  and  $(a, x) \geq 0$ ,  $x \in L_1$ . Then  $0 \leq (a, x) = (a, A^T y) = (Aa, y)$  for all  $y \in \mathbf{R}^m$ . This is possible only if  $Aa = 0$ ,  $a \in L$ , which contradicts the condition  $(a, c) < 0$ . Therefore  $c \in L_1$ .  $\square$

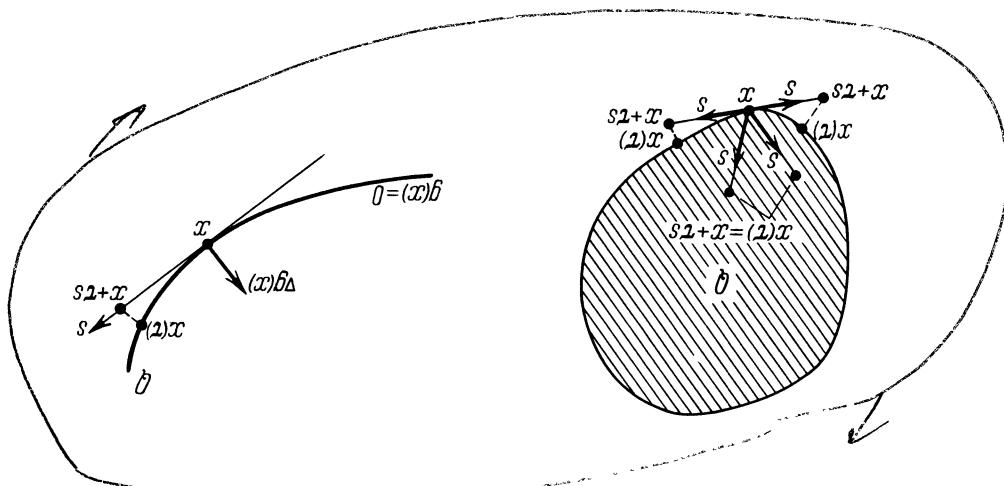


Fig. 33 Tangent vectors.

**PROOF** of Theorem 2. Let  $s$  be a tangent vector to the set  $Q = \{x: g_i(x) = 0, i = 1, \dots, m\}$  at  $x^*$ . Then we can find  $x(\tau)$  such that  $g_i(x(\tau)) = 0, i = 1, \dots, m$ ,  $\|x^* + \tau s - x(\tau)\| = o(\tau)$ . Hence

$$f(x(\tau)) = f(x^* + \tau s + o(\tau)) = f(x^*) + \tau (\nabla f(x^*), s) + o(\tau).$$

Since  $f(x(\tau)) \geq f(x^*)$  for sufficiently small  $\tau$ , then  $(\nabla f(x^*), s) \geq 0$ . By Lyusternik's theorem,  $(s, \nabla g_i(x^*)) = 0, i = 1, \dots, m$ . Using Lemma 1, we obtain  $\nabla f(x^*) = \sum_{i=1}^m \mu_i \nabla g_i(x^*)$ , where the  $\mu_i$  are some scalars. Setting  $y_i^* = -\mu_i, i = 1, \dots, m$ , we arrive at (2).  $\square$

2. *Proof Based on Elimination of Variables.* If  $\nabla g_i(x^*)$  are linearly independent, then the matrix  $g'(x^*)$ , whose rows are  $\nabla g_1(x^*), \dots, \nabla g_m(x^*)$ , has rank  $m$ . Therefore we can find  $m$  components of the vector  $x$  (we denote the set of them by  $I$ ), such that the matrix with elements  $\partial g_j(x^*)/\partial x_i, j = 1, \dots, m, i \in I$ , has an inverse. We write the vector  $x \in \mathbf{R}^n$  in the form  $\{u, v\}$ , where  $u \in \mathbf{R}^m$  are the components of  $x$  with indices in  $I$ ,  $v \in \mathbf{R}^{n-m}$  are the remaining components. Then the matrix  $g'_u(u^*, v^*)$  (where  $g(u, v) = g(x)$ ,  $x^* = \{u^*, v^*\}$ ) has an inverse. Consider the equality  $g(u, v) = 0$ . Since  $g(u^*, v^*) = 0$ ,  $g$  is continuously differentiable in a neighborhood of  $\{u^*, v^*\}$  and the matrix  $g'_u(u^*, v^*)$  is nonsingular, then by the implicit function theorem (Theorem 2 of Section 2.3) we can find a differentiable function  $u(v)$  in a neighborhood of  $v^*$  such that

$$u(v^*) = u^*, \quad g(u(v), v) = 0$$

and

$$u'(v) = -[g'_u(u(v), v)]^{-1} g'_v(u(v), v).$$

Next we consider the function  $\phi(v) = f(u(v), v)$ , where  $f(u, v) = f(x)$ . The function  $\phi(v)$  attains a local unconstrained minimum at  $v^*$ . Indeed, for any  $v$  close to  $v^*$ ,  $g(u(v), v) = 0$ , i.e., the point  $x = (u(v), v)$  is admissible, and therefore

$$\phi(v^*) = f(u(v^*), v^*) = f(u^*, v^*) = f(x^*) \leq f(x) = f(u(v), v) = \phi(v).$$

Hence  $\nabla \phi(v^*) = 0$ . By the chain rule for differentiating a composite function,

$$\nabla \phi(v) = u'(v)^T f'_u(u(v), v)^T + f'_v(u(v), v)^T.$$

Thus

$$0 = \nabla \phi(v^*) = -g'_v(u^*, v^*)^T [g'_u(u^*, v^*)^T]^{-1} f'_u(u^*, v^*)^T + f'_v(u^*, v^*)^T.$$

Let

$$[g'_u(u^*, v^*)^T]^{-1} f'_u(u^*, v^*)^T = -y^*. \quad (6)$$

Then

$$f'_u(u^*, v^*)^T + g'_u(u^*, v^*)^T y^* = 0, \quad f'_v(u^*, v^*)^T + g'_v(u^*, v^*)^T y^* = 0,$$

which is equivalent to equality (2).  $\square$

This proof is based on the idea of reducing a constrained problem to an unconstrained minimum problem by means of elimination of variables. To be precise, the variables  $x \in \mathbf{R}^n$  are divided into two groups,  $u \in \mathbf{R}^m$ , and  $v \in \mathbf{R}^{n-m}$ ; from the equalities  $g(x) = 0$  we express one group in terms of the other:  $u = u(v)$  and consider the unconstrained minimum problem for  $\phi(v) = f(u(v), v)$ . The necessary minimum condition for it ( $\nabla \phi(v^*) = 0$ ) gives an extremum condition for the initial problem. In this case formula (6) gives an explicit expression for the Lagrange multipliers.

**3. Proof Based on Penalty Functions.** Let  $U = \{x: \|x - x^*\| \leq \varepsilon\}$ , where  $\varepsilon > 0$  is such that  $f, g_i$  are continuously differentiable on  $U$  and  $x^*$  is the global minimum point on  $Q \cap U$ .

Consider the problem

$$\min_{x \in U} f_k(x), \quad f_k(x) = f(x) + \frac{1}{2} K \sum_{i=1}^m g_i^2(x) + \frac{1}{2} \|x - x^*\|^2, \quad (7)$$

where  $K$  is some parameter. By the continuity of  $f_k(x)$ , problem (7) has a solution  $x^k$ . Therefore

$$\begin{aligned} f_k(x^k) &\leq f_k(x^*), \\ f_k(x) + \frac{1}{2} K \sum_{i=1}^m g_i^2(x^k) + \frac{1}{2} \|x^k - x^*\|^2 &\leq f(x^*), \\ \sum_{i=1}^m g_i^2(x^k) &\leq \frac{2}{K} (f(x^*) - f(x^k) - \frac{1}{2} \|x^k - x^*\|^2). \end{aligned}$$

The quantity on the right-hand side tends to 0 as  $K \rightarrow \infty$  (since  $\|x^k - x^*\| \leq \varepsilon$ ), therefore  $g(x^k) \rightarrow 0$ . Let a sequence  $x^{k_i} \rightarrow \tilde{x} \in U$ . Then  $g(\tilde{x}) = 0$ ,  $f(\tilde{x}) + \|\tilde{x} - x^*\|^2/2 \leq f(x^*)$ ; on the other hand, since  $x^*$  is a minimum point on  $Q$ , then  $f(x^*) \leq f(\tilde{x})$ . Hence  $\tilde{x} = x^*$ . Since every limit point for  $x^k$  coincides with  $x^*$ , one has  $x^k \rightarrow x^*$  as  $K \rightarrow \infty$ . Therefore for sufficiently large  $K > 0$ , the point  $x^k$  lies inside  $U$ . Thus, the minimum condition for it takes the form  $\nabla f_k(x^k) = 0$  i.e.,

$\lambda = 0$

$$\nabla f(x^k) + K \sum_{i=1}^m g_i(x^k) \nabla g_i(x^k) + x^k - x^* = 0. \quad (8)$$

Let

$$y_0^k = \frac{1}{\sqrt{1 + K^2 \sum_{i=1}^m g_i^2(x^k)}}, \quad y_i^k = \frac{K g_i(x^k)}{\sqrt{1 + K^2 \sum_{i=1}^m g_i^2(x^k)}}, \quad i = 1, \dots, m.$$

Equality (8) can be written in the form

$$y_0^k \nabla f(x^k) + \sum_{i=1}^m y_i^k \nabla g_i(x^k) + (x^k - x^*) y_0^k = 0. \quad (9)$$

We have  $\sum_{i=0}^m (y_i^k)^2 = 1$  for all  $k$ , and therefore we can find a sequence  $k_j \rightarrow \infty$  such that

$$\cancel{y_i^k} \rightarrow y_i^*, \quad i = 0, \dots, m, \quad \sum_{i=0}^m (y_i^*)^2 = 1.$$

$\cancel{y_i^k} \rightarrow y_i^*$

Passing to the limit in (9) yields (1).  $\square$

In our proof we have used the same idea as in the preceding proof, viz. the necessary extremum condition in the unconstrained problem is invoked to obtain a necessary condition in the constrained problem. However, the method for reducing the one problem to the other is very different: we construct the sequence ( $K \rightarrow \infty$ ) of unconstrained minimization problems that differ by an increasing “penalty” for violating the constraints (the term  $(\frac{1}{2})K \sum_{i=1}^m g_i^2(x)$  in  $f_k(x)$ ) the solutions of which tend in the limit to solutions of the initial constrained minimization problem.

The last proof is the simplest in terms of the tools used, e.g., Lyusternik's theorem, the implicit function theorem, or any similar assertions have not been invoked.

### Exercises

1. Consider problem (A) in  $\mathbf{R}^2$  with  $f(x) = x_2$ ,  $g_1(x) = (x_1 - 1)^2 + x_2^2 - 1$ ,  $g_2(x) = (x_1 + 1)^2 + x_2^2 - 1$ , (Fig. 34). Show that  $x^* = \{0, 0\}$  is not a regular minimum point and (2) is not satisfied at this point.
2. Check that if  $g_i(x) = (a^i, x) - b_i$ ,  $i = 1, \dots, m$ , then (2) coincides with (8) of Section 7.1.
3. Show that if  $Q$  is convex, then the tangent cone  $S$  is convex and coincides with  $T$ , that is the closure of the cone generated by the feasible directions (see the proof of Theorem 3 in Section 7.1).

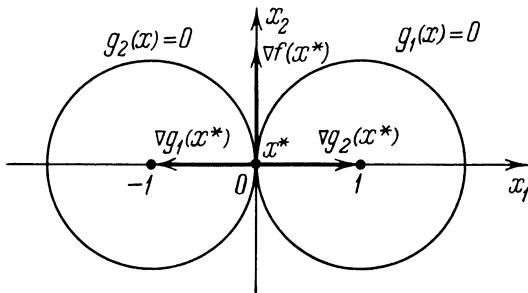


Fig. 34 The problem with a nonregular minimum.

4. Check that, for example, in Exercise 1 Lyusternik's theorem does not apply for \$x^\* = 0\$ and (5) does not hold.
5. Show that if the point \$x^\*\$ is a locally unique minimum point, then one need not include the term \$\|x - x^\*\|^2/2\$ in the function \$f\_k(x)\$ (see proof 3).
6. Check that if the minimum is regular, then for the \$Kg\_i(x^k)\$ (see proof 3) there exists a limit as \$K \rightarrow \infty\$: \$Kg\_i(x^k) \rightarrow y\_i^\*/y\_0^\*\$.

### 8.1.2 Second-order Minimum Conditions

Theorem 4 (necessary second-order condition). Let \$x^\*\$ be a regular minimum point in problem (A), let \$f(x)\$ and \$g\_i(x)\$ be twice continuously differentiable in a neighborhood of \$x^\*\$, and let \$y\_i^\*\$, \$i = 1, \dots, m\$, be Lagrange multipliers. Then

$$(L''_{xx}(x^*, y^*)s, s) = \left[ (\nabla^2 f(x^*) + \sum_{i=1}^m y_i^* \nabla^2 g_i(x^*))s, s \right] \geq 0 \quad (10)$$

for all

$$s \in S = \{s: (\nabla g_i(x^*), s) = 0, i = 1, \dots, m\}.$$

In other words, the matrix \$L''\_{xx}(x^\*, y^\*)\$ is nonnegative definite on the tangent space \$S\$ (see (5)).

**PROOF.** Let \$s \in S\$. By Lyusternik's theorem, there are admissible \$x(\tau)\$ such that \$\|x^\* + \tau s - x(\tau)\| = o(\tau)\$. Then, using (4), we obtain

$$\begin{aligned}
f(x^*) &\leq f(x(\tau)) = L(x(\tau), y^*) \\
&= L(x^*, y^*) + (L'_x(x^*, y^*), x(\tau) - x^*) \\
&\quad + (L''_{xx}(x^*, y^*)(x(\tau) - x^*), x(\tau) - x^*)/2 + o(\tau^2) \\
&= f(x^*) + (\tau^2/2)(L''_{xx}(x^*, y^*)s, s) + o(\tau^2),
\end{aligned}$$

yielding  $(L''_{xx}(x^*, y^*)s, s) \geq 0$ .  $\square$

The more general necessary extremum condition  $L''_{xx}(x^*, y^*) \geq 0$  that seems natural at first glance is in fact false (see Exercise 8).

Before proceeding to consider sufficient extremum conditions, we formulate some auxiliary results concerning matrices of special form, which we shall be using in our later discussion.

**LEMMA 2.** Let  $A$  be a symmetric  $n \times n$  matrix, let  $C$  be an  $m \times n$  matrix of rank  $m$ , and let  $(Ax, x) > 0$  for all  $x \neq 0$  such that  $Cx = 0$ . Then the block matrix

$$B = \begin{pmatrix} A & C^T \\ C & 0 \end{pmatrix} \quad (11)$$

of dimension  $(m+n) \times (m+n)$  is invertible.  $\square$

**LEMMA 3.** Under the conditions of Lemma 2, there is a  $K_0 > 0$ ,  $\alpha > 0$ , such that  $A + KCTC \geq \alpha I$  for  $K \geq K_0$ .  $\square$

**LEMMA 4.** Under the conditions of Lemma 2, for sufficiently large  $K$  one has

$$\begin{aligned}
\|(A + KCTC)^{-1}\| &\leq \alpha_1/K, \quad \|C(A + KCTC)^{-1}\| \leq \alpha_2/K, \\
\|I - KC(A + KCTC)^{-1}C^T\| &\leq \alpha_3/K,
\end{aligned}$$

$\vdash J^{-1} C^T \|$

where  $\alpha_i$  are constants.  $\square$

**LEMMA 5.** Under the conditions of Lemma 2, for the matrix

$$B_k = \begin{pmatrix} A & C^T \\ C & -\frac{1}{K}I \end{pmatrix} \quad (12)$$

$\checkmark$

for sufficiently large  $K$ ,  $B_k^{-1}$  exists and  $\|B_k^{-1}\| \leq \gamma$ .  $\square$

Now let us return to formulating extremum conditions.

**THEOREM 5** (sufficient second-order condition). Let  $g_i(x^*) = 0$ ,  $i = 1, \dots, m$ , let the functions  $f(x)$  and  $g_i(x)$  be twice continuously differentiable in a neighborhood of  $x^*$ , and let  $\nabla g_i(x^*)$ ,  $i = 1, \dots, m$ , be linearly independent. Furthermore, let the necessary minimum condition (4) be satisfied and let

$$\sqrt{L''_{xx}(x^*, y^*)s, s} > 0 \quad (13)$$

for all  $s$  such that  $(\nabla g_i(x^*), s) = 0$ ,  $i = 1, \dots, m$ . Then  $x^*$  is a local minimum point in problem (A).

In other words, if the necessary first-order extremum condition holds at  $x^*$  and the matrix  $L''_{xx}(x^*, y^*)$  is positive definite on the tangent subspace  $S$ , then  $x^*$  is a minimum point. We say that a point  $x^*$  at which the conditions of Theorem 5 are satisfied is a nonsingular point.

**PROOF.** Introduce the function

$$M(x, y, K) = f(x) + (y, g(x)) + \frac{(K/2)\|g(x)\|^2}{=} = L(x, y) + \frac{(K/2)\|g(x)\|^2}{=} \quad (14)$$

where  $K > 0$  is some parameter. Then

$$M'_x(x^*, y^*, K) = L'_x(x^*, y^*) = 0, \\ M''_{xx}(x^*, y^*) = L''_{xx}(x^*, y^*) + K g'(x^*)^T g'(x^*).$$

For the matrices  $A = L''_{xx}(x^*, y^*)$  and  $C = g'(x^*)$  Lemma 3 is applicable. Hence for sufficiently large  $K > 0$ ,

$$M''_{xx}(x^*, y^*, K) > 0. \quad (15)$$

Thus, the sufficient local minimum condition for  $M(x, y^*, K)$  is satisfied (Theorem 4 of Section 1.2), i.e.,  $M(x, y^*, K) \geq M(x^*, y^*, K)$  for all  $x$  close enough to  $x^*$ . But for  $x \in Q$  (i.e., for admissible  $x$ ) we have  $M(x, y^*, K) = f(x)$ , i.e.,  $f(x) \geq f(x^*)$  for  $x \in Q$  in a neighborhood of  $x^*$ .  $\square$

The function  $M(x, y, K)$  in (14) is called the *augmented* Lagrange function. It plays an important role in constrained optimization theory. Let us examine some of its properties. First, it is different from the usual Lagrangian (3) by the “penalty” term  $(K/2)\|g(x)\|^2$  and coincides with it for  $K = 0$ :  $M(x, y, 0) = L(x, y)$ . Furthermore, if  $x \in Q$ , then

$$M(x, y, K) = L(x, y) = f(x) \quad \text{and} \quad M'_x(x, y, K) = L'_x(x, y),$$

while

$$M'_y(x, y, K) = L'_y(x, y) = g(x).$$

Hence the necessary first-order minimum condition has the form analogous to (4):

$$M'_x(x^*, y^*, K) = 0, \quad M'_y(x^*, y^*, K) = 0, \quad (16)$$

where the Lagrange multipliers  $y^*$  are the same as in (4).

However, the  $M(x, y, K)$  and  $L(x, y)$  begin to differ with respect to the second-order conditions. As was shown in proving Theorem 5, if  $x^*$  is a nonsingular minimum point in the constrained problem (A), then  $x^*$  is a nonsingular unconstrained minimum point of  $M(x, y^*, K)$  for sufficiently large  $K$ . For the ordinary Lagrangian the analogue is false, i.e., the point  $x^*$  is a stationary point of  $L(x, y^*)$ , but not necessarily a minimum point (see Exercise 8). This property of the augmented Lagrangian can be employed in constructing efficient optimization methods (Section 8.2).

### Exercises

7. Show that in the problem  $\min f(x)$ ,  $Ax = b$ , the necessary minimum conditions are:

$$\nabla f(x^*) + A^T y^* = 0, \quad (\nabla^2 f(x^*) s, s) \geq 0,$$

for all  $s$  such that  $As = 0$ .

8. The problem

$$\min f(x), \quad x \in \mathbf{R}^2, \quad g(x) = 0, \quad f(x) = x_1^2 - x_2^2, \quad g(x) = x_2,$$

has the solution  $x^* = \{0, 0\}$ , with  $y^* = 0$ . Verify that the matrix  $L''_{xx}(x^*, y^*)$  is indefinite.

9. Let  $A, D$  be symmetric matrices of dimensions  $n \times n$  and  $m \times m$ , let  $C$  be an  $m \times n$  matrix, and let  $B = \begin{pmatrix} A & C^T \\ C & D \end{pmatrix}$  be an  $(n+m) \times (n+m)$  matrix. Prove now that the condition  $B \geq 0$  is equivalent to the conditions  $A \geq 0$ ,  $CA^+C^T - D \geq 0$ , and  $B > 0$  is equivalent to the conditions  $A > 0$ ,  $CA^{-1}C^T - D > 0$  (a generalization of Sylvester's criterion to the matrix case).

10. Prove that under the conditions of Theorem 5 of this Section,  $x^* = \underset{x \in S}{\operatorname{argmin}} L(x, y^*)$ .

#### 8.1.3 The Usage of Extremum Conditions

In standard courses in calculus, the study of constrained minimization problems ends with a derivation of extremum conditions. The view is that such conditions make it possible to find the solution. This is not, however, true. The Langrange multipliers rule determines the system of equations (4)

in  $x^*$ ,  $y^*$ . These equations are nonlinear (with the exception of quadratic  $f(x)$  and linear  $g_i(x)$ ), and the solution to these equations cannot be found, as a rule, in the explicit form. The examples given in textbooks to illustrate the possibility of solving problems through Lagrange multipliers are exceptions from the rule, specially selected to prove the point (such as those we give in Exercise 11).

The real role that extremum conditions play is different (cf. the analogous remarks in Section 1.2), viz. (1) in constructing numerical methods for finding a solution, (2) after the solution has been found, they help evaluate the uniqueness, stability, and other properties of this solution (see below), and (3) they define the requirements for which it is convenient to analyze the problem, e.g., investigate the convergence of the methods.

The reader will find in the sequel many examples of this usage of extremum conditions.

### Exercise

**11.** Find the solutions of the following problems, using Lagrange multipliers, and prove the optimality of your results, using the sufficient extremum conditions:

- (a)  $\min \sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i = 1;$
- (b)  $\min \sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2 = 1;$
- (c)  $\min (Ax, x), \|x\| = 1;$
- (d)  $\min \|x\|^2, (Ax, x) = 1.$

ANSWER: (a)  $x_i^* = 1/n, i = 1, \dots, n$ ; (b)  $x_i^* = -1/\sqrt{n}, i = 1, \dots, n$ ; (c)  $x^* = e^1$ , the normalized eigenvector corresponding to the largest eigenvalue of the matrix  $A$ ; (d)  $x^* = \lambda_n^{-1/2} e^n$ ,  $e^n$  being the normalized eigenvector corresponding to the largest eigenvalue  $\lambda_n$  of the matrix  $A$ , the solution exists for  $\lambda_n > 0$ .

#### 8.1.4 Existence, Uniqueness and Stability of a Solution

The question of the existence of a solution is resolved again by Theorem 4 of Section 7.1; in this case the specific features of problem (A) are of no consequence.

With regard to uniqueness of a solution, it is usually impossible to use the theorem on uniqueness of the minimum of a strictly convex function on a convex set for problem (A), because a set  $Q$  defined by nonlinear equality constraints is not convex (with the exception of nonsingular cases) (see Exercise 12). In this case, however, one can introduce *a posteriori* uniqueness conditions.

**THEOREM 6.** A nonsingular minimum point is locally unique.

Indeed, in proving Theorem 5 we obtained that a nonsingular solution  $x^*$  of problem (A) is a nonsingular unconstrained minimum point of  $M(x, y^*, K)$ . Hence we can find an  $\ell > 0$  such that

$$M(x, y^*, K) - M(x^*, y^*, K) \geq \ell \|x - x^*\|^2$$

in some neighborhood of  $x^*$  (see (2) in Section 1.3). Since  $f(x) = M(x, y^*, K)$  for  $x \in Q$ , then

$$f(x) - f(x^*) \geq \ell \|x - x^*\|^2 \quad (17)$$

for  $x \in Q$  in a neighborhood of  $x^*$ .  $\square$

The next result on uniqueness of Lagrange multipliers, derived from the definition of a regular point, is immediate.

**THEOREM 7.** For a regular minimum the Lagrange multipliers are uniquely determined.  $\square$

To analyze the stability, we consider first the stability of a solution with respect to constraint perturbations. Along with the initial problem (A), we introduce the “perturbed” problem

$$\begin{aligned} \min f(x) , \\ g_i(x) = \varepsilon_i , \quad i = 1, \dots, m , \end{aligned} \quad (18)$$

where  $\varepsilon = (\varepsilon_1, \dots, \varepsilon_m) \in \mathbf{R}^m$  is some vector. Let  $x_\varepsilon$  denote the solution to this problem (if it exists) and let  $\phi(\varepsilon) = f(x_\varepsilon)$ . We are interested in the case where  $x_\varepsilon \rightarrow x^*$  as  $\varepsilon \rightarrow Q$  ( $x^*$  is the solution of (A)), as well as in estimation of the proximity of  $x_\varepsilon$  to  $x^*$  and the behavior of  $\phi(\varepsilon)$  for small  $\varepsilon$ .

**THEOREM 8.** Let  $x^*$  be a nonsingular solution to problem (A). Then for sufficiently small  $\|\varepsilon\|$  there exists an  $x_\varepsilon$ ,

$$\|x_\varepsilon - x^*\| = O(\varepsilon) , \quad \nabla \phi(0) = -y^* . \quad (19)$$

**PROOF.** Let  $z = \{x, y\} \in \mathbf{R}^{n+m}$ ,  $x \in \mathbf{R}^n$ ,  $y \in \mathbf{R}^m$ ,  $R(z) = \{L'_x(x, y), L'_y(x, y)\}$ . Then the system of equations (4) can be written in the form

$$R(z) = 0 . \quad (20)$$

Obviously,  $R(z^*) = 0$ , where  $z^* = \{x^*, y^*\}$ ,  $x^*$  is the solution of problem (A),  $y^*$  are the corresponding Lagrange multipliers. Let us compute  $R'(z^*)$ . We have

$$\mathcal{L}_x R'(z^*) = \begin{pmatrix} A & C^T \\ C & 0 \end{pmatrix}, \quad A = L''_{xx}(x^*, y^*), \quad C = L''_{y\neq j}(x^*, y^*) = g'(x^*). \quad (21)$$

It follows from Lemma 2 that  $R'(z^*)$  is nonsingular. By Theorem 3 of Section 2.3 the system

$$R(z) = a \quad (22)$$

has a solution  $z_a$  for sufficiently small  $\|a\|$ , and

$$z_a = z^* - [R'(z^*)]^{-1}a + o(a). \quad (23)$$

Take  $a = \{0, \varepsilon\}$ ,  $a \in \mathbf{R}^{n+m}$ ,  $\varepsilon \in \mathbf{R}^m$ . In this case (22) is equivalent to the system

$$\nabla f(x) + g'(x)^T y = 0, \quad g(x) = \varepsilon, \quad (24)$$

and it has a solution  $z_\varepsilon = \{x_\varepsilon, y_\varepsilon\}$  for sufficiently small  $\varepsilon$ . Therefore, the point  $x_\varepsilon$  (1) satisfy the constraints of problem (A), (2) by the continuity of  $\nabla g_i(x)$  and the regularity of  $x^*$ , the gradients  $\nabla g_i(x_\varepsilon)$  are also linearly independent for sufficiently small  $\|\varepsilon\|$ , (3) at  $x_\varepsilon$  we have, by (24), the necessary minimum condition in problem (18) with Lagrange multipliers  $y_\varepsilon$ , and (4) by the continuity of the first and second derivatives and the linear independence of the  $\nabla g_i(x^*)$  we have the condition  $L''_{xx}(x_\varepsilon, y_\varepsilon) > 0$  on the subspace  $S_\varepsilon = \{s: (\nabla g_i(x_\varepsilon), s) = 0, i = 1, \dots, m\}$ . Thus, at  $x_\varepsilon$  we have the sufficient second-order extremum condition, i.e.,  $x_\varepsilon$  is a solution of problem (18). It follows from (23) that  $\|z_\varepsilon - z^*\| \leq \alpha \|a\|$ ,  $\alpha$  being some constant, and hence  $\|x_\varepsilon - x^*\| \leq \alpha \|\varepsilon\|$ . Finally,

$$\begin{aligned} f(x) &\geq f(x) + (y^*, g(x)) = L(x, y^*) \geq L(x^*, y^*) \\ &= f(x^*) + (y^*, g(x^*)) = f(x^*), \end{aligned}$$

$$\begin{aligned} \phi(\varepsilon) &= f(x_\varepsilon) = f(x^*) + (\nabla f(x^*), x_\varepsilon - x^*) + o(\|x_\varepsilon - x^*\|) \\ &= f(x^*) - (g'(x^*)^T y^*, x_\varepsilon - x^*) + o(\varepsilon) \\ &= f(x^*) - (g(x_\varepsilon) - g(x^*), y^*) + o(\varepsilon) = f(x^*) - (y^*, \varepsilon) + o(\varepsilon). \end{aligned}$$

Therefore,  $\nabla \phi(0) = -y^*$ .  $\square$

Theorem 8 implies the stability of a nonsingular minimum toward perturbations in the constraints. In particular, if for a sequence  $x^k$  in a neighborhood of the nonsingular minimum  $x^*$  one has

$$\lim_{k \rightarrow \infty} g_i(x^k) = 0, \quad i = 1, \dots, m, \quad \lim_{k \rightarrow \infty} f(x^k) = f(x^*),$$

then this sequence converges to  $x^*$ .

Stability toward perturbations of the objective function can be analyzed in a similar way. To illustrate, we give a typical example.

**THEOREM 9.** Let  $x_\varepsilon$  be a solution of the problem  $[f(x) + \varepsilon h(x)]$ ,  $g_i(x) = 0$ ,  $i = 1, \dots, m$ , where  $\varepsilon \in \mathbf{R}^1$ ,  $h(x)$  is a twice continuously differentiable function in a neighborhood of  $x^*$ ,  $x^*$  being a nonsingular solution of problem (A). Then for small  $|\varepsilon|$ ,  $x_\varepsilon$  exists and  $x_\varepsilon \rightarrow x^*$  as  $\varepsilon \rightarrow 0$ .  $\square$

The assumption concerning the nonsingularity of a minimum is essential in Theorems 8 and 9. For example, in the problem  $\min x$ ,  $g(x) = x^2 = 0$ ,  $x \in \mathbf{R}^1$ , its solution  $x^* = 0$  being non-regular. For the perturbed problem with the constraints  $g(x) = \varepsilon$  for  $\varepsilon < 0$  a solution does not exist (the admissible set is empty), whereas for  $\varepsilon > 0$ ,  $x_\varepsilon = \sqrt{\varepsilon}$  and bound (19) is violated.

### Exercise

12. Let  $g: \mathbf{R}^n \rightarrow \mathbf{R}^1$  be a strictly convex function. Prove that if the set  $\{x: g(x) = 0\}$  contains more than one point, then it is nonconvex.

## 8.2 MINIMIZATION METHODS

### 8.2.1 Classification of the Methods

Methods for solving constrained optimization problems are numerous and diverse. They can be classified by form as well as content.

As before, it is possible to define the methods of zero, first and second orders, depending on the order of the derivatives involved. We shall deal basically with the first-order methods (in which the gradients  $\nabla f(x)$  and  $\nabla g_i(x)$  are computed) and second-order methods (where it is required to know  $\nabla^2 f(x)$  and  $\nabla^2 g_i(x)$ ). Furthermore, the methods are classified as primal (in which the iterations are executed on the space of primal variables  $x$ ) and dual (which essentially use the dual variables  $y$ ). In many methods, an auxiliary problem is solved at each step, and for the convenience of computations the methods should be classified by the type of auxiliary problem involved. This may be an unconstrained minimization problem, or a problem for minimizing a linear or a quadratic function under linear constraints, etc.

Finally, the ideas on which the methods are based vary extensively, e.g., elimination of variables, linearization, penalty functions, ordinary and augmented Lagrangians, among others. We shall examine the most important methods—in terms of concepts as well as computations.

### 8.2.2 The Linearization Method

In this method, the objective function and the constraints are linearized in each iteration. Since the problem of minimizing a linear function under linear constraints may not have a solution, a quadratic term is added to the function (cf. the similar technique used in constructing the gradient method (3) of Section 1.4). We have thus arrived at a method in which the recursive approximation  $x^{k+1}$  is a solution of the following auxiliary problem:

$$\begin{aligned} \min & [(\nabla f(x^k), x - x^k) + (2\gamma)^{-1} \|x - x^k\|^2], \\ & g_i(x^k) + (\nabla g_i(x^k), x - x^k) = 0, \quad i = 1, \dots, m, \end{aligned} \tag{1}$$

where  $\gamma > 0$  is some parameter. We have already come across this kind of problem more than once (Chapter 7). On the one hand, method (1) can be written as the gradient-projection method for linearized constraints:

$$\begin{aligned} x^{k+1} &= P_{Q_k}(x^k - \gamma \nabla f(x^k)), \\ Q_k &= \{x: g(x^k) + g'(x^k)(x - x^k) = 0\}. \end{aligned} \tag{2}$$

On the other hand, the solution of the system of linear equations

$$\begin{aligned} (1/\gamma)(x - x^k) + g'(x^k)^T y &= -\nabla f(x^k), \\ g'(x^k)(x - x^k) &= -g(x^k) \end{aligned} \tag{3}$$

is the vector  $\{x^{k+1}, y^{k+1}\} \in \mathbf{R}^{n+m}$ , the first components of which coincide with  $x^{k+1}$ . Thus, to find  $x^{k+1}$  it suffices to solve the system of linear equations (3) (of dimension  $n+m$ ). The resulting vector  $y^{k+1}$  is, as we shall see later, a bound for the Lagrange multipliers  $y^*$ .

**THEOREM 1.** Let  $x^*$  be a nonsingular minimum point, and let  $\nabla^2 f(x)$ ,  $\nabla^2 g_i(x)$  satisfy a Lipschitz condition in a neighborhood of  $x^*$ . Then we can find a  $\bar{\gamma} > 0$  such that for  $0 < \gamma < \bar{\gamma}$  method (1) is well-defined and converges locally to  $x^*$  with the rate of geometric progression.

**PROOF.** First, by the linear independence of the  $\nabla g_i(x^*)$ ,  $i = 1, \dots, m$ , the vectors  $\nabla g_i(x^k)$ ,  $i = 1, \dots, m$ , are linearly independent as well for  $x^k$  sufficiently close to  $x^*$ , and therefore  $Q_k \neq \emptyset$  and method (2) is well-defined, i.e., the point  $x^{k+1}$  exists. By (2), this method can be written in the form (see (6) in Section 7.2)

$$\begin{aligned} x^{k+1} &= (I - C_k^+ C_k)(x^k - \gamma \nabla f(x^k)) + C_k^+(C_k x^k - g(x^k)), \\ C_k &= g'(x^k). \end{aligned} \quad (4)$$

Since

$$C_k = C + g''(x^*)(x^k - x) + o(x^k - x^*),$$

where

$$\begin{aligned} C &= g'(x^*), \quad \nabla f(x^k) = \nabla f(x^*) + \nabla^2 f(x^*)(x^k - x^*) + o(x^k - x^*), \\ \nabla f(x^*) + C^T y^* &= 0, \end{aligned}$$

while  $C^+ = C^T(CCT)^{-1}$ , then for  $x^k$  sufficiently close to  $x^*$ , we have

$$\begin{aligned} x^{k+1} - x^* &= D(x^k - x^*) + o(x^k - x^*), \\ D &= (I - C^T(CCT)^{-1}C)(I - \gamma A), \quad A = L''_{xx}(x^*, y^*). \end{aligned} \quad (5)$$

Let us show that for sufficiently small  $\gamma > 0$  one has  $\rho(D) < 1$ , where  $\rho(D)$  is the spectral radius of  $D$ . Indeed, consider the iterations  $u^{k+1} = Du^k$  for arbitrary  $u^0 \in \mathbf{R}^n$ . Since  $Du = P_S(I - \gamma A)u$ ,  $S = \{x: Cx = 0\}$ , then all the  $u^k$  belong to  $S$ ,  $k \geq 1$ . Hence for  $k \geq 1$ ,

$$\begin{aligned} \|(I - \gamma A)u^k\|^2 &= \|u^k\|^2 - 2\gamma(Au^k, u^k) + \gamma^2\|Au^k\|^2 \\ &\leq \|u^k\|^2 - 2\gamma\ell\|u^k\|^2 + \gamma^2\|A\|^2\|u^k\|^2, \end{aligned}$$

since  $(Au, u) \geq \ell\|u\|^2$ ,  $\ell > 0$ , for  $u \in S$  by the nonsingularity of  $x^*$ . Thus,  $\|(I - \gamma A)u^k\| \leq q\|u^k\|$ ,  $q < 1$ , for small  $\gamma > 0$ . But  $\|I - C^T(CCT)^{-1}C\| \leq 1$  since  $I - C^T(CCT)^{-1}C$  is the projection operator; see (6) of Section 5.1. Therefore,  $\|u^{k+1}\| \leq q\|u^k\|$ ,  $q < 1$ ,  $k \geq 1$ , i.e.,  $u^k \rightarrow 0$ , which is equivalent to  $\rho(D) < 1$  (Corollary of Lemma 1 in Section 2.1). Now applying Theorem 1 of Section 2.1 to (5) yields the desired result.  $\square$

Theorem 1 is typical in many respects.

First, only the local convergence of the method is proved. This is quite natural since in problems with nonlinear equality constraints the admissible

set is generally nonconvex, as was noted earlier. Hence no global result is expected in such problems.

Second, the basic tool for proving Theorem 1 is Theorem 1 of Section 2.1. Furthermore, the iterative process is considered for the primal and dual variables ( $x$  and  $y$ ) at the same time, and the theorem is applied in the space  $\mathbb{R}^{n+m}$ .

Third, it is assumed that the minimum is nonsingular. If it is singular, one can prove sometimes the convergence of some or other first-order method, but it is impossible to guarantee the convergence with the rate of geometric progression.

Fourth, in Theorem 1 no explicit expression is given for the parameters ( $\bar{\gamma}$ , the progression ratio, the size of the region of convergence). In principle, such bounds can be written out, but the expressions would be cumbersome, or they would contain *a priori* unknown parameters (e.g.,  $y^*$ ). Hence we shall usually limit ourselves to assertions similar to Theorem 1, which evaluate qualitatively the behavior of the method.

### 8.2.3 Dual Methods

The linearization method does not contain explicitly the Lagrangian or the dual variables (although, as the reader saw, it does give approximations for the Lagrange multipliers). In the method described below (also known as the Arrow-Hurwicz method) the primal and dual variables are equipotent:

$$\begin{aligned} \mathcal{L}_x & x^{k+1} = x^k - \gamma L'_x(x^k, y^k) = x^k - \gamma(\nabla f(x^k) + g'(x^k)^T y^k), \\ & y^{k+1} = y^k + \gamma L'_y(x^k, y^k) = y^k + \gamma g(x^k). \end{aligned} \quad (6)$$

In other words, one makes a step of the gradient method for minimizing the Lagrangian in  $x$  and simultaneously a step for maximizing the Lagrangian in  $y$ . Such a method can hardly be expected to converge in a general situation. Indeed, as we noted, the function  $L(x, y^*)$  does not need to attain a minimum in  $x$  at  $x^*$ . However, if  $y^0 = y^*$ , then (6) turns into the gradient method for  $L(x, y^*)$ , which, as we know from Theorem 1 of Section 6.2, does not converge to a stationary point that is not a minimum point. Hence Theorem 2 below contains the additional condition that the matrix  $L''_{xx}(x^*, y^*)$  be positive definite.

$\mathcal{L}_x$  **THEOREM 2.** Let  $x^*$  be a nonsingular minimum point, let  $L''_{xx}(x^*, y^*) > 0$  and let the second derivatives  $\nabla^2 f(x)$ ,  $\nabla^2 g_i(x)$  satisfy a Lipschitz condition in a neighborhood of  $x^*$ . Then we can find a  $\bar{\gamma} > 0$  such that for  $0 < \gamma < \bar{\gamma}$  method (6) converges locally to  $x^*$ ,  $y^*$  with the rate of geometric progression.

**PROOF.** In the notation  $z^k = \{x^k - x^*, y^k - y^*\}$  the method can be written as

$$\begin{aligned} z^{k+1} &= Dz^k + o(z^k), \\ D = I - \gamma B, \quad B &= \begin{pmatrix} A & C^T \\ -C & 0 \end{pmatrix}, \quad A = L''_{xx}(x^*, y^*), \quad C = g'(x^*). \end{aligned} \tag{7}$$

We now show that the matrix  $-B$  is stable. This is equivalent to the fact that for the system  $\dot{z} = -Bz$  one has  $z(t) \rightarrow 0$  as  $t \rightarrow \infty$  and for any  $z(0)$  (see Lemma 3 of Section 2.1). In the variables  $x, y$  the system takes on the form:  $\dot{x} = -Ax - C^T y$ ,  $\dot{y} = Cx$ . Take  $\rho(t) = (\|x(t)\|^2 + \|y(t)\|^2)/2$ . Then

$$\rho' = (\dot{x}, x) + (\dot{y}, y) = -(Ax, x) - (C^T y, x) + (Cx, y) = -(Ax, x) \leq -\alpha \|x\|^2$$

since  $A > 0$ . Hence  $\rho$  decreases monotonically, as  $\rho \rightarrow 0$ , and therefore  $x(t) \rightarrow 0$  as  $t \rightarrow \infty$ . Since  $x(t)$  is a solution of a linear differential equation, then  $x(t) \rightarrow 0$  implies  $\dot{x}(t) \rightarrow 0$ . Thus,  $Cy = -\dot{x} - Ax \rightarrow 0$  as  $t \rightarrow \infty$ . Since  $\text{rank } C = m$ , this yields  $y \rightarrow 0$ . Thus  $z(t) \rightarrow 0$ . By Lemma 5 of Section 2.1, the spectral radius of  $D = I - \gamma B$  is less than 1 for sufficiently small  $\gamma > 0$ . Applying Theorem 1 of Section 2.1 for (7) yields the required result.  $\square$

One can modify method (6), making a complete minimization of the Lagrangian in  $x$  instead of one step of the gradient method:

$$x^{k+1} = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} L(x, y^k), \quad y^{k+1} = y^k + \gamma g(x^{k+1}). \tag{8}$$

It turns out that under the conditions of Theorem 2 the method is well-defined for  $y^0$  sufficiently close to  $y^*$  (i.e.,  $x^{k+1}$  exists), and for (8) the assertions of Theorem 2 hold true. Thus, passage to the more laborious method (8) (in which one needs to solve an unconstrained minimization problem at each step) does not alter the qualitative evaluation of the behavior of method (6), even though the progression ratio may have decreased.

#### 8.2.4 The Augmented Lagrangian Method

Methods of the type (6), (8), in which the usual Lagrangian is replaced by the augmented Lagrangian possess significantly better properties. We consider first an analogue of method (6):

$$\begin{aligned} x^{k+1} &= x^k - \gamma M'_x(x^k, y^k, K) \\ &= x^k - \gamma(\nabla f(x^k) + g'(x^k)^T y^k + K g'(x^k)^T g(x^k)), \\ y^{k+1} &= y^k + \gamma M'_y(x^k, y^k, K) = y^k + \gamma g(x^k). \end{aligned} \tag{9}$$

**THEOREM 3.** Let  $x^*$  be a nonsingular minimum point, and let  $\nabla^2 f(x)$  and  $\nabla^2 g_i(x)$  satisfy a Lipschitz condition in a neighborhood of  $x^*$ . Then for a sufficiently large  $K$  and  $0 < \gamma < \bar{\gamma}$ , method (9) converges locally to  $x^*$ ,  $y^*$  with the rate of geometric progression.

**PROOF.** Problem (A) is equivalent to the problem

$$\begin{aligned} \min [f(x) + (K/2) \|g(x)\|^2], \\ g(x) = 0, \end{aligned} \tag{10}$$

for which the Lagrangian coincides with  $M(x, y, K)$  and method (6) turns into (9). However, by virtue of (15) of Section 8.1,  $M''_{xx}(x^*, y^*, K) > 0$ , and therefore Theorem 2 on convergence of method (6) and thereby (9) is applicable.  $\square$

Thus, method (9) converges for any nonsingular minimum point without the additional assumptions required for methods (6) and (8), based on the use of a Lagrangian. Using the augmented Lagrangian, a sufficiently high rate of convergence can be achieved. This applies to the analog of method (8) as well:

$$x^{k+1} = \underset{x}{\operatorname{argmin}} M(x, y^k, K), \quad y^{k+1} = y^k + K g(x^{k+1}). \tag{11}$$

**THEOREM 4.** Let the conditions of Theorem 3 be satisfied. Then for every  $y^0$  sufficiently close to  $y^*$ , we can find a  $K_0$  such that for  $K > K_0$  method (11) converges to  $x^*$ ,  $y^*$  with the rate of geometric progression, with the ratio  $q = O(1/K)$ .

**PROOF.** Set  $\phi(x) = M(x, y^*, K)$ ,  $\psi_k(x) = (y^k - y^*, g(x))$ . Then  $M(x, y^k, K) = \phi(x) + \psi_k(x)$ . By (15) and (16) of Section 8.1,  $x^*$  is a nonsingular unconstrained minimum point of  $\phi(x)$ . By Theorem 6 of Section 3.1, if  $\|y^k - y^*\|$  is sufficiently small, then there exists a  $x^{k+1}$ , a local minimum point of  $\phi(x) + \psi_k(x)$  in a neighborhood of  $x^*$ , and

$$x^{k+1} - x^* = -[\nabla^2 \phi(x^*)]^{-1} \nabla \psi_k(x^*) + o(y^k - y^*).$$

Since

$$\nabla^2 \phi(x^*) = A + KC^T C, \quad g'(x) = C, \quad A = L''_{xx}(x^*, y^*),$$

$$\nabla \psi_k(x^*) = C^T (y^k - y^*),$$

by Lemma 4 of Section 8.1 we have

$$\begin{aligned}\|x^{k+1} - x^*\| &\leq \|(A + KC^T C)^{-1} C^T\| \|y^k - y^*\| + o(y^k - y^*) \\ &\leq (\alpha_1/K) \|y^k - y^*\| + o(y^k - y^*).\end{aligned}\tag{12}$$

We apply this lemma again:

$$\begin{aligned}y^{k+1} &= y^k + Kg(x^{k+1}) = y^k + KC(x^{k+1} - x^*) + o(x^{k+1} - x^*) \\ &= y^k - KC(A + KC^T C)^{-1} C^T (y^k - y^*) + o(y^k - y^*), \\ \|y^{k+1} - y^k\| &\leq \|I - KC(A + KC^T C)^{-1} C^T\| \|y^k - y^*\| + o(y^k - y^*) \\ &\leq (\alpha_3/K) \|y^k - y^*\| + o(y^k - y^*).\end{aligned}$$

Thus, for sufficiently small  $y^0 - y^*$ , one has  $y^k \rightarrow y^*$  with the rate of geometric progression, with ratio  $q = O(1/K)$ , and from (12) it then follows that  $x^k \rightarrow x^*$  with the same rate.  $\square$

A more accurate accounting for the remainder terms in the proof of Theorem 4 yields a stronger assertion.

**THEOREM 5.** The results of Theorem 4 hold for any  $y^0$  ( $K_0$  depending on  $\|y^0 - y^*\|$ ).  $\square$

Therefore, method (11) is superior to method (8) in some respects. It does not call for a good initial approximation in  $y$ , which is of importance since such an approximation is usually unknown. Also, it converges under the minimal assumptions: it suffices for the functions to be smooth and the minimum to be nonsingular. In terms of computations, method (11) is not more complicated than method (8). It does not require a special choice of the step size  $\gamma$  (however, there arises the problem of the appropriate choice of  $K$ ). Moreover, most importantly, the convergence rate of method (11) can be increased by the choice of  $K$ .

Of course, one should not overestimate the advantages of method (11). In the formulations of Theorems 4 and 5 the words “converges locally” are missing, but this does not mean that the method will allow us to find the global solution of the problem—it means that we have a difficult situation with the auxiliary problem (11). In determining  $x^{k+1}$ , it is assumed implicitly that  $x^{k+1}$  is an unconstrained minimum point of  $M(x, y^k, K)$  close to  $x^*$  (such a point exists); however finding such a point is still a problem. Furthermore, if  $K$  is very large, then the convergence rate of the iteration

*r*(*K*) in (11) increases, each iteration becoming, however, more laborious. The fact is the problem of minimizing  $M(x, y, K)$  becomes ill-posed.

The question of a compromise choice of  $K$  taking into account the foregoing conditions is a difficult one; it has not yet been resolved. In practical computations one can change  $K$  in each iteration, making it depend on the computational results.

### 8.2.5 The Penalty Function Method

To solve problem (A), we shall apply the same idea of reduction to a sequence of unconstrained minimization problems as that in proving the Lagrange multipliers rule using the penalty functions (see (7) in Section 8.1):

$$x^k = \underset{x \in Q_0}{\operatorname{argmin}} f_k(x), \quad f_k(x) = f(x) + \frac{1}{2}K_k \|g(x)\|^2, \quad K_k \rightarrow \infty, \quad (13)$$

where  $Q_0$  is some bounded set of localization of the minimum introduced as to the unconstrained minimization problem have a solution. We shall prove the convergence of this method under the weakest assumptions (we do not require even that  $f$  and  $g_i$  be differentiable).

**THEOREM 6.** Let problem (A) have solutions, and let  $X^*$  denote the set of all these solutions. Let  $f$  and  $g_i$  be continuous, and let  $Q_0$  be closed and bounded,  $Q_0 \cap X^* \neq \emptyset$ . Then every limit point of method (13) (by  $x^k$  we mean the global minimum of  $f_k(x)$  on  $Q_0$ ) is a global minimum for problem (A).

**PROOF.** Method (13) is well-defined since  $f_k(x)$  is continuous and  $Q_0$  is closed and bounded. Hence by Theorem 4 of Section 7.1 the points  $x^k$  exist. Since  $x^k \in Q_0$ , there exists at least one limit point  $\tilde{x}$  for the sequence  $x^k$ . Let  $x^* \in Q_0 \cap X^*$ . Then by the definition of (13), we have  $f_k(x^k) \leq f_k(x^*) = f_k(x^*)$ , yielding  $\|g(x^k)\|^2 < (2/K_k)(f(x^*) - f(x^k))$ , and passing to the limit as  $k \rightarrow \infty$  we obtain  $g(\tilde{x}) = 0$ , i.e.,  $\tilde{x}$  is an admissible point. On the other hand,  $f(x^k) \leq f(x^*) - (\frac{1}{2})K_k \|g(x^k)\|^2 \leq f(x^*)$ , and hence  $f(\tilde{x}) \leq f(x^*)$ . Then  $\tilde{x}$  is solution to problem (A).  $\square$

For a nonsingular minimum one can also estimate the convergence rate of (13).

**THEOREM 7.** Let  $x^*$  be a nonsingular minimum point and let  $\nabla^2 f(x)$ ,  $\nabla^2 g_i(x)$  satisfy a Lipschitz condition in a neighborhood of  $x^*$ . Then  $x^k \rightarrow x^*$  as  $K_k \rightarrow \infty$ , where

$$\|x^k - x^*\| = O(1/K_k), \quad K_k g(x^k) \rightarrow y^*. \quad \square \quad (14)$$

Theorem 7 does not, however, imply that by the appropriate choice of  $K_k$  we can get an arbitrarily high convergence rate: as was noted before, as  $K_k$  grows the condition numbers of the auxiliary problems diminish and their solution becomes more and more difficult. In the penalty-function method one must take  $K_k \rightarrow \infty$  (otherwise it will not converge), and this is the major disadvantage of this method versus the augmented Lagrangian method, where one can do without increasing  $K$ . For practical calculations, the  $K_k$  are increased as to make the  $x^{k-1}$  the initial approximation for finding the  $x^k$  (when  $K_k$  grows fast, the region of convergence of unconstrained minimization methods for  $f_k(x)$  is reduced, and  $x^{k-1}$  may not fall into this region). Note that the penalty-function method coincides with a variant of the augmented Lagrangian method, in which the dual variables are not updated ( $y^k \equiv 0$ ) (see Exercise 2). Generally, in the penalty-function method, the dual variables are not used at all, although they can be found using (14). In fact, the augmented Lagrangian method can be viewed as a variation of the penalty-function method, in which the information about the Lagrange multipliers has been used systematically. This variation proved to be significantly more efficient than the initial method. The only possible advantage of the penalty-function method is its great universality: it converges under very weak conditions (see Theorem 6).

### Exercises

1. Prove Theorem 7 following the same scheme as in Theorem 4, writing (13) in the form  $x^k = \underset{x}{\operatorname{argmin}} M(x, 0, K_k)$ .
2. Examine the method  $x^k = \underset{x}{\operatorname{argmin}} M(x, y^0, K_k)$  for some constant  $y_0$  and show that all the results hold with respect to the penalty-function method (in which  $y^0 = 0$ ). Prove that the convergence rate is the higher, the closer  $y^0$  is to  $y^*$ .

### 8.2.6 The Reduced Gradient Method

The elimination of variables applied in the second method of proving the Lagrange multipliers rule (Section 8.1) leads us to the reduced gradient method. Let  $x = \{u, v\}$ ,  $u \in \mathbf{R}^m$ ,  $v \in \mathbf{R}^{n-m}$ , be the division of the variables into two groups, where  $u$  can be found in terms of  $v$  from the equation  $g(x) = g(u, v) = 0$ . We construct the gradient method of unconstrained minimization of the function  $\phi(v) = f(u(v), v)$ , where  $f(u, v) = f(x)$ , i.e.,

$$\begin{aligned} v^{k+1} &= v^k - \gamma \nabla \phi(v^k), \\ \nabla \phi(v^k) &= -g'_v(u^k, v^k)^T [g'_u(u^k, v^k)^T]^{-1} f'_u(u^k, v^k)^T + f'_v(u^k, v^k)^T, \end{aligned} \quad (15)$$

where  $u^k$  is the solution of the equation  $g(u, v^k) = 0$ . Thus in method (15)

there is no need to find the dependence of  $u(v)$  in the explicit form; it suffices to solve the equation  $g(u, v^k) = 0$  for fixed  $v^k$ .

**THEOREM 8.** Let  $x^*$  be a nonsingular minimum point and let  $\nabla^2 f, \nabla^2 g_i$  satisfy a Lipschitz condition in a neighborhood of  $x^*$ . Then we can find a  $\bar{\gamma} > 0$  such that for  $0 < \gamma < \bar{\gamma}$  method (15) converges locally to  $\{u^*, v^*\} = x^*$  with the rate of geometric progression.

The proof consists in verifying the nonsingularity of the unconstrained minimum point  $v^*$  of  $\phi(v)$  and applying Theorem 4 of Section 1.4.  $\square$

Let us examine separately the special case of linear constraints. Suppose the constraints have the form

$$A_1 u + A_2 v = b, \quad (16)$$

where  $A_1$  is a nonsingular  $m \times m$  matrix,  $A_2$  is a  $m \times (n-m)$  matrix,  $b \in \mathbf{R}^m$ , and the objective function is separable in  $u$  and  $v$ :

$$f(x) = f_1(u) + f_2(v). \quad (17)$$

Method (15) takes on the form

$$\begin{aligned} v^{k+1} &= v^k - \gamma(\nabla f_2(v^k) - A_2^T (A_1^T)^{-1} \nabla f_1(u^k)), \\ u^{k+1} &= A_1^{-1}(b - A_2 v^{k+1}), \end{aligned} \quad (18)$$

and converges globally if  $f_1(u)$  and  $f_2(v)$  are strongly convex and smooth. One can also guarantee global convergence, provided  $f_2(v)$  is strongly convex, whereas  $f_1(u)$  has a sufficiently small second derivative.

### 8.2.7 Newton's Method

To solve the system of equations (4) in Section 8.1, one can apply Newton's method, i.e., the new approximation  $x^{k+1}, y^{k+1}$  is sought as a solution of the system of linearized equations

$$\begin{aligned} L''_{xx}(x^k, y^k)(x-x^k) &\neq L''_{yx}(x^k, y^k)(y-y^k) = -L'_x(x^k, y^k), \\ L''_{xy}(x^k, y^k)(x-x^k) &= -L''_y(x^k, y^k) \end{aligned} \quad (19)$$

(since  $L''_{yy}(x, y) \equiv 0$ ) or, in more detailed notation,

$$\begin{aligned} \sqrt{(20)} \left[ \nabla^2 f(x^k) + \sum_{i=1}^m y_i^k \nabla^2 g_i(x^k) \right] (x-x^k) + g'(x^k)^T (y-y^k) &= -\nabla f(x^k) - g'(x^k)^T y^k, \\ g'(x^k)(x-x^k) &= -g(x^k). \end{aligned}$$

**THEOREM 9.** Let  $x^*$  be a nonsingular minimum point and let  $\nabla^2 f$  and  $\nabla^2 g_i$  satisfy a Lipschitz condition in a neighborhood of  $x^*$ . Then method (20) converges locally to  $x^*, y^*$  with quadratic rate.

**PROOF.** As before (see the proof of Theorem 8 in Section 8.1), we introduce

$$z = \{x, y\} \in \mathbf{R}^{n+m}, \quad z^k = \{x^k, y^k\}, \quad z^* = \{x^*, y^*\}.$$

Then (4) in Section 8.1 can be written in the form

$$R(z) = 0, \quad R(z) = \{L'_x(x, y), L'_y(x, y)\}, \quad (21)$$

$$R: \mathbf{R}^{n+m} \rightarrow \mathbf{R}^{n+m},$$

while (20) becomes Newton's method for solving this equation, i.e.,  $z^{k+1}$  is a solution of the system of linearized equations

$$R'(z^k)(z - z^k) = -R(z^k). \quad (22)$$

Since  $R(z^*) = 0$ ,  $R'(z)$  satisfies a Lipschitz condition in a neighborhood of  $z^*$  and the matrix  $R'(z^*)$  is nonsingular (see (21) and Lemma 2 of Section 8.1), then the general result on convergence of Newton's method is applicable (Theorem 3 of Section 1.5), which yields the required assertion.  $\square$

Newton's method (20) has the same advantages and disadvantages as Newton's method for unconstrained minimization: it converges rapidly but requires laborious computation of the second derivatives and a good initial approximation. Such an approximation is especially difficult to find for the dual variables.

It is possible to put Newton's method in a different form. We compose a quadratic approximation (in  $x$ ) of the Lagrangian for a fixed value  $y^k$  and seek the minimum of this approximation under linear constraints. For  $y^{k+1}$  we take  $y^k + u^k$ , where  $u^k$  are the Lagrange multipliers for the auxiliary problem

$$\begin{aligned} \min & [(L'_x(x^k, y^k), x - x^k) + (L''_{xx}(x^k, y^k)(x - x^k), x - x^k)/2], \\ & g(x^k) + g'(x^k)(x - x^k) = 0. \end{aligned} \quad (23)$$

It is not hard to prove that methods (20) and (23) are equivalent. (Cf. the different forms for writing the linearization method (1)-(3).)

### 8.2.8 Other Quadratically Convergent Methods

As noted earlier (see Exercise 10 in Section 8.1), we have

$$x^* = \underset{x \in S}{\operatorname{argmin}} L(x, y^*) , \quad S = \{x: g'(x^*)(x - x^*) = 0\} \quad (24)$$

for a nonsingular minimum point  $x^*$ . Hence it is natural to construct a method in which, at the  $k$ th step, one is seeking the minimum of  $L(x, y^k)$  on the subspace formed by linearization of the constraint  $g(x) = 0$  at  $x^k$ :

$$\begin{aligned} x^{k+1} &= \underset{x \in Q_k}{\operatorname{argmin}} L(x, y^k) , \\ Q_k &= \{x: g(x^k) + g'(x^k)(x - x^k) = 0\} , \\ y^{k+1} &= y^k + u^k , \end{aligned} \quad (25)$$

where  $u^k$  are the Lagrange multipliers in problem (25). This method is very close to method (23), the only difference being that on  $Q_k$  one seeks the minimum of  $L(x, y^k)$  rather than the minimum of its quadratic approximation. Therefore, in this method there is no need to calculate the second derivatives: minimization on the subspace can be done by any efficient first-order method, say the conjugate gradient method (6) of Section 7.3. Of course, method (25) has the same properties as method (23).

Other methods similar to method (23) can be suggested, e.g., quasi-Newton methods in which the auxiliary problem

$$\underset{x \in Q_k}{\operatorname{min}} [L'_x(x^k, y^k)(x - x^k) + \frac{1}{2}(H_k(x - x^k), x - x^k)] \quad (26)$$

is solved, where the matrix  $H_k$  is an approximation for  $L''_{xx}(x^k, y^k)$  constructed from the previous values of the gradients.

## 8.3 HOW TO HANDLE POSSIBLE COMPLICATIONS

In the preceding section we considered an ideal situation: we ignored the presence of noise, limited ourselves to local results and the case of nonsingular minimum, etc. Let us see what risk this ideal situation involves.

### 8.3.1 A Global Minimum

All the assertions of Sections 8.1 and 8.2 were “local,” which, as was noted, is due to nonconvexity of problem (A) (cf. Exercise 12 of Section 8.1). If the initial problem (A) has no local minima, even then the system of equations (4) of Section 8.1 (i.e., the notation for the rule of Lagrange multipliers) does not, as a rule, a unique solution. For instance, in the problem

$$\min_{\|x\|^2=1} (Ax, x) , \quad (1)$$

where  $A$  is a symmetric  $n \times n$  matrix, there are no local minima, yet every eigenvector of  $A$  satisfies the necessary extremum conditions, which in this particular case take on the form  $Ax + yx = 0$ ,  $y \in \mathbf{R}^1$ . It is no surprise that results on convergence of minimization methods were mainly local, i.e., required a good initial approximation of the solution. An exception is the penalty-function method: its convergence is “global” (Theorem 6 of Section 8.2). Of course, this method implies the problem of seeking the global minimum in auxiliary problems (see Section 6.2). One may specify a reasonable penalty coefficient and, using local methods of Section 8.2, make descent from the “suspect” points obtained in the global minimization process.

To summarize, the problem of seeking the global solution to problem (A) is really complicated, and there is no universal method for solving it.

### 8.3.2 Noise

The stability theorems in Section 8.2 make one expect that small noise in the computation of the functions and their gradients would not be a problem in the nonsingular case. Indeed, it can be shown that for a non-singular minimum those methods, for sufficiently small absolute noise, lead to a neighborhood of the solution, the size of which is the smaller, the lower the noise level.

Here is a typical result. Consider the “perturbed” linearization method:

$$x^{k+1} = \underset{x \in Q_k}{\operatorname{argmin}} \left[ (\nabla \tilde{f}(x^k), x - x^k) + \frac{1}{2\gamma} \|x - x^k\|^2 \right], \quad (2)$$

$$\tilde{Q}_k = \{x : \tilde{g}(x^k) + \nabla \tilde{g}(x^k)(x - x^k) = 0\},$$

$\nabla \tilde{g}$

where for all  $x$  in some neighborhood  $U$  of the point  $x^*$ ,

$$\|\nabla f(x) - \nabla f(x^*)\| \leq \varepsilon_1, \quad \|\tilde{g}(x) - g(x)\| \leq \varepsilon_2,$$

$$\|\nabla g(x) - \nabla g(x^*)\| \leq \varepsilon_3.$$

**THEOREM 1.** Under the conditions of Theorem 1 of Section 8.2, we can find an  $\varepsilon_0 > 0$  such that for every  $\varepsilon > 0$  there are  $\delta_i > 0$ ,  $i = 1, 2, 3$ , such that in method (2) one has  $\|x^k - x^*\| \leq \varepsilon$  for all sufficiently large  $k$  if  $\|x_0 - x^*\| \leq \varepsilon_0$ ,  $\varepsilon_i < \delta_i$ ,  $i = 1, 2, 3$ .

Let us sketch the proof. As in Section 8.2, we can show that

$$x^{k+1} - x^* = D(x^k - x^*) + o(x^k - x^*) + r^k,$$

where

$$\|r^k\| = O(\varepsilon_1 + \varepsilon_2 + \varepsilon_3), \quad \rho(D) < 1.$$

Let  $U > I$  be a solution of the matrix equation  $D^T UD = U - I$  (Lemma 2 of Section 2.1),  $v_k = (U(x^k - x^*), x^k - x^*)$ . Then (cf. the proof of Theorem 2 of Section 2.1) for sufficiently small  $v_0$  one has  $v_{k+1} \leftarrow qv_k + \alpha_k$ , where  $q < 1$ ,  $\alpha_k = O(\epsilon_1 + \epsilon_2 + \epsilon_3)$ , thus proving the theorem.  $\square$

Completely analogous assertions apply as well to the other methods in Section 8.2.

However, if the noise is not small enough, a complete “breakdown” of the methods can occur. For example, suppose that in calculating the  $g_i(x)$  a systematic error  $\epsilon_i$  is made,  $i = 1, \dots, m$ . This results in the situation where the problem with constraints of the form  $g_i(x) = \epsilon_i$ ,  $i = 1, \dots, m$ , is being solved. It may happen that such a system has no solution even if the initial problem has a nonsingular solution  $x^*$ . In this situation, any method can yield a meaningless result (see Section 8.3.4 below). We wish to point out that it was not characteristic of unconstrained minimization problems (see Chapter 4).

We shall not go into a detailed discussion of problems with relative noise. In this case much depends on the meaning one puts into this term. If we assume that

$$\|r(x)\| \leq \alpha \|x - x^*\|, \quad (3)$$

where  $r(x)$  are all possible errors in computing the gradients and the functions, then it can be shown that for a nonsingular minimum the convergence of the methods in Section 8.2 continues to hold for sufficiently small  $\alpha$ . But if we assume that the errors satisfy conditions such as

$$\|\tilde{\nabla}f(x) - \nabla f(x)\| \leq \alpha \|\nabla f(x)\|, \quad (4)$$

then this situation is essentially equivalent to the case with absolute noise since  $\nabla f(x^*) \neq 0$ .

Finally, the case of absolute noise is somewhat special. Consider, for example, Arrow-Hurwicz method (see (6) of Section 8.2):

$$\begin{aligned} x^{k+1} &= x^k - \gamma_k (\tilde{\nabla}f(x^k) + g'(x^k)^T y^k), \\ y^{k+1} &= y^k + \gamma_k g(x^k), \end{aligned} \quad (5)$$

where  $\tilde{\nabla}f(x^k) = \nabla f(x^k) + \xi^k$ , and the  $\xi^k$  is independent centered random noise (for simplicity we assume that  $g_i(x)$  and  $\nabla g_i(x)$  are known exactly). Since the  $\xi^k$  not generally bounded, there exists a nonzero probability of “ejection” of a point from the region of convergence. As a result, using Theorem 6 of Section 2.2, one can prove only the following: if  $\sum_{k=0}^{\infty} \gamma_k = \infty$ ,  $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$ , then under the conditions of Theorem 2 of Section 8.2 me-

thod (5) converges to a solution with probability  $1 - \delta$ , where  $\delta$  is the smaller, the smaller the noise variance and the more precise the initial approximation. Similar results hold also in the presence of noise in calculating the  $g_i(x)$  and  $\nabla g_i(x)$ , as well as for several other methods in Section 8.2.

### 8.3.3 A Singular Minimum

The majority of the results of Section 8.2 were proved under the assumption that the minimum was nonsingular. This assumption may need to be modified for two reasons:

(i) at the point  $x^*$  the gradients of the constraints may not be linearly independent. For such problems, generally, the Lagrange multipliers rule does not hold and the tangent subspace theorem is inapplicable. Hence all methods based on the ordinary and the augmented Lagrange function, as well as on constraint linearization, may become ineffective. Let us consider an example in  $\mathbb{R}^2$ , which we have come across before (Exercise 1 of Section 8.1 and Fig. 34):

$$\min x_2 ,$$

$$g_1(x) = (x_1 - 1)^2 + x_2^2 - 1 = 0 , \quad (6)$$

$$g_2(x) = (x_1 + 1)^2 + x_2^2 - 1 = 0 .$$

The solution is  $x^* = \{0, 0\}$ , and  $\nabla g_1(x^*) = \{-2, 0\}$  and  $\nabla g_2(x^*) = \{2, 0\}$  are linearly dependent. Then at  $x^0 = \{\varepsilon, 0\}$  linearization of the constraints leads to the set

$$Q_0 = \{x: g_1(x^0) + (\nabla g_1(x^0), x - x_0) = 0, g_2(x^0) + (\nabla g_2(x^0), x - x_0) = 0\} ,$$
VV

which is empty for any  $\varepsilon \neq 0$ . Thus, the linearization method is inapplicable for an initial approximation close to  $x^*$ . Newton's method becomes meaningless for the very same reason (see its formula in (23) of Section 8.2). Finally, all the dual methods for the given problem cannot converge since the Lagrange multipliers do not exist for it. Apparently, one can construct methods based on extremum conditions for the nonregular case (Theorem 1 of Section 8.1). However, this approach has not been investigated so far.

(ii) a nonsingular minimum point may be regular, but only the weaker necessary condition (10) of Section 8.1 can be satisfied instead of the sufficient extremum condition (13) of Section 8.1. This situation resembles in many ways the singular minimum case for unconstrained problems (see Section 1 of Chapter 6). Thus, for problems with linear constraints and convex  $f(x)$  one can prove the convergence of the linearization method and of the methods based on the augmented Lagrange function; however, it is

impossible to assert in this case the convergence rate of geometric progression. At the same time, a method using the ordinary Lagrangian does not necessarily converge. Consider the simplest example in  $\mathbf{R}^1$ :  $\min f(x)$ ,  $g(x) = 0$ , where  $f(x) = x$ ,  $g(x) = x$ . Then  $x^* = 0$  is the solution,  $y^* = -1$ ,  $L(x, y^*) \equiv 0$ , so that  $L''_{xx}(x, y^*) \equiv 0$ , and the conditions of Theorem 2 of Section 8.2 do not hold. In this case, method (6) of Section 8.2 takes on the form  $x^{k+1} = x^k - \gamma(y^k + 1)$ ,  $y^{k+1} = y^k + \gamma x^k$ . Hence for  $\rho_k = (x^k - x^*)^2 + (y^k - y^*)^2$ , we obtain  $\rho_{k+1} = \rho_k(1 + \gamma^2)$ , i.e.,  $\rho_k \rightarrow \infty$  for any  $\gamma \neq 0$ .

The least sensitive to all forms of singularity is the penalty-function method. Under conditions for its convergence (Theorem 6 of Section 8.6), there is no condition for the minimum to be nonsingular (or even regular). However, singularity may slow down the convergence of the penalty-function method. To see this is the case, one need not construct special examples —it suffices to observe that this method is related to the regularization method. Indeed, the problem  $\min [f(x) + K\|g(x)\|^2]$  is equivalent to the problem  $\min [\varepsilon f(x) + \|g(x)\|^2]$  with  $\varepsilon = 1/K$ , which can be viewed to be the problem  $\min \|g(x)\|^2$  regularized by means of the function  $f(x)$ . However the regularization method, as we know (see Section 6.1), may converge very slowly in the singular case.

### 8.3.4 Incompatibility of Constraints

It is possible that problem (A) is ill-posed, viz. the set of admissible points is empty. The reason for this is manifold. Frequently, in engineering and economics problems the initial requirements imposed on the “object” are too demanding, and therefore, conflicting. Sometimes, the fault lies with errors in the available data, or errors in characteristics of the objects, and the like. Finally, if problem (A) arises as an auxiliary problem while solving more complex problems, the constraints can be incompatible because the approximation used is no longer valid. For example, if the linearization method of Section 8.2 is used and the initial approximations are bad, the resulting auxiliary problems may have no solution.

A moot question is how the various minimization methods will fare under such conditions. Some of the methods will simply not apply. For instance, the linearization method and Newton's method will not be correctly defined because of contradictory constraint conditions. Other methods may apply, but will not converge. Indeed, the dual methods will clearly diverge since there is no point at which  $g(x)$  vanishes. The penalty-function method will behave better. For example, if the constraints are linear and  $f(x)$  is convex, we do have convergence to a pseudosolution: the minimum of  $\|g(x)\|^2$  at which  $f(x)$  is minimal.

## CHAPTER 9

### A GENERAL PROBLEM OF MATHEMATICAL PROGRAMMING

In this chapter we shall investigate the general problem of mathematical programming:

$$\begin{aligned} \min f(x) , \quad & x \in \mathbf{R}^n , \\ g_i(x) \leq 0 , \quad & i = 1, \dots, r , \\ g_i(x) = 0 , \quad & i = r+1, \dots, m , \\ x \in Q , \end{aligned} \tag{A}$$

where  $Q \subset \mathbf{R}^n$  is a “simple” set (cf. Chapter 7), and  $g_i: Q \rightarrow \mathbf{R}^1$ ,  $i = 1, \dots, m$ . We say that points at which all the constraints are satisfied are *admissible*. Special cases of problem (A) were investigated in Chapter 7 ( $r = m = 0$ ) and Chapter 8 ( $r = 0$ ,  $Q = \mathbf{R}^n$ ). We shall analyze two basic classes of problems (A): nonlinear programming ( $f(x)$ ,  $g_i(x)$  are differentiable,  $Q = \mathbf{R}^n$ ) and convex programming ( $f(x)$ ,  $g_i(x)$ ,  $i = 1, \dots, r$ , are convex,  $r = m$ , and  $Q$  being convex).

#### 9.1 THE THEORY OF CONVEX PROGRAMMING

##### 9.1.1 Convex Analysis: Fundamentals

In addition to the fundamentals of convex analysis given in Section 5.1 we shall need here the following.

We considered previously convex functions defined on the entire space  $\mathbf{R}^n$ . It makes sense to extend the class of convex functions by dropping the last condition. Let  $Q \subset \mathbf{R}^n$  be some set, and let the scalar function  $f(x)$  be defined on  $Q$  and be not defined outside  $Q$ . We call  $Q$  the domain of definition of  $f(x)$  and denote it by  $D(f)$ . Sometimes, the notation  $\text{dom } f$  and the term "effective domain" are used. We call a function  $f(x)$  *convex* on  $D(f)$  if for any  $x \in D(f)$ ,  $y \in D(f)$ ,  $0 \leq \lambda \leq 1$ , one has  $\lambda x + (1-\lambda)y \in D(f)$  and

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

It follows from the definition that if  $f(x)$  is convex, then  $D(f)$  is a convex set.

It is convenient to approach the notion of a function proper from a different standpoint. Assume that the function can take on not only finite values but the value  $+\infty$  as well, which obeys the standard rules of arithmetic operations and inequalities:

$$\begin{aligned} \alpha < \infty, \quad \infty + \alpha &= \infty, \quad \alpha \cdot \infty = \infty \quad (\alpha > 0), \\ \infty \leq \infty, \quad \infty + \infty &= \infty, \quad \max \{\alpha, \infty\} = \infty, \end{aligned} \quad (2)$$

for all  $\alpha \in \mathbf{R}^1$ . The expressions  $\infty - \infty$ ,  $\infty/\infty$  are not defined, but we do set  $0 \cdot \infty = 0$ . Taking this into account, we complete the definition of the convex function  $f(x)$  given on  $D(f) \subset \mathbf{R}^n$  to all the  $\mathbf{R}^n$ , setting

$$f(x) = +\infty \quad \text{for } x \notin D(f). \quad (3)$$

Then inequality (1) remains valid for all  $x, y \in \mathbf{R}^n$ . Now, by a *convex* function we mean a function  $f(x)$  taking on values from  $\mathbf{R}^1 \cup \{+\infty\}$ , defined on the entire space  $\mathbf{R}^n$  and satisfying (1) for all  $x, y \in \mathbf{R}^n$ ,  $0 \leq \lambda \leq 1$ . In this case

$$D(f) = \{x: f(x) < \infty\}, \quad (4)$$

and, in addition, we assume without specifying it each time that  $D(f)$  is nonempty (sometimes the term *eigenvalued* convex functions is used). Similarly, we say that a function  $f(x)$  with values in  $\mathbf{R}^1 \cup \{-\infty\}$  is *concave* if  $-f(x)$  is convex, and write  $D(f) = \{x: f(x) > -\infty\}$ . To illustrate, we consider the following four scalar functions (Figs. 35 (a)-(d)):

$$f(x) = \begin{cases} 0, & |x| \leq 1, \\ \infty, & |x| > 1; \end{cases} \quad (5)$$

$$f(x) = \begin{cases} 0, & |x| < 1, \\ 1, & |x| = 1, \\ \infty, & |x| > 1; \end{cases} \quad (6)$$

$\leftarrow$  proper

$$f(x) = \begin{cases} 1 - \sqrt{1-x^2}, & |x| \leq 1, \\ \infty, & |x| > 1; \end{cases} \quad (7)$$

$$f(x) = \begin{cases} x^2(1-x^2)^{-1}, & |x| < 1, \\ \infty, & |x| \geq 1. \end{cases} \quad (8)$$

It is clear that they are convex in the sense of the above definition; furthermore, for (5)-(7)  $D(f) = [-1, 1]$  and for (8)  $D(f) = (-1, 1)$ . The usual properties of finite convex functions remain valid with marginal refinements.

**LEMMA 1.** (a) If  $f(x)$  is convex, then  $\alpha f(x)$  is convex for  $\alpha > 0$ ,  $D(\alpha f) = D(f)$  and the sets  $\{x: f(x) \leq \alpha\}$ ,  $\{x: f(x) < \alpha\}$ ,  $\{x: f(x) < \infty\}$  are convex for  $\alpha \in \mathbf{R}^1$ .

(b) If  $f_1(x), f_2(x)$  are convex and  $D(f_1) \cap D(f_2) \neq \emptyset$ , then the functions  $f(x) = f_1(x) + f_2(x)$  and  $f(x) = \max \{f_1(x), f_2(x)\}$  are convex and  $D(f) = D(f_1) \cap D(f_2)$ .  $\square$

A convex function taking on the value  $+\infty$  at any point is discontinuous at this point, and hence Lemma 3 of Section 5.1 on continuity of a finite convex function does not extend to the case  $D(f) \neq \mathbf{R}^n$ . However, we have the following lemma.

**LEMMA 2.** A convex function is continuous at every point of  $D(f)^0$  (here and in the sequel  $D(f)^0$  denotes the interior of  $D(f)$ ).  $\square$

The behavior of a convex function  $f(x)$  on the boundary of  $D(f)$  can vary. Example (6) shows that the conditions  $x^k \rightarrow x^*, x^k \in D(f), x^* \in D(f)$  do not necessarily imply  $f(x^k) \rightarrow f(x^*)$ .

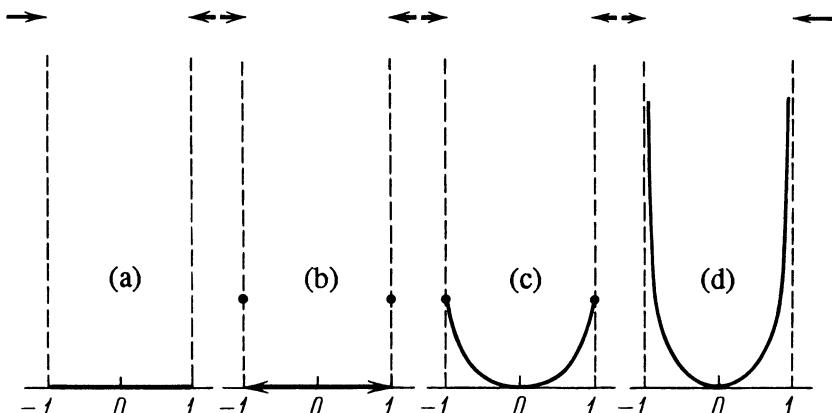


Fig. 35 Convex functions with bounded domain of definition.

As earlier, we say that a vector  $a \in \mathbf{R}^n$  is the subgradient of a convex function  $f(x)$  at  $x \in \mathbf{R}^n$  and write  $\partial f(x)$  if

$$f(x + y) \geq f(x) + (a, y) \quad (9)$$

for all  $y \in \mathbf{R}^n$ . Then, if  $D(f) = \mathbf{R}^n$ , this definition coincides with (10) in Section 5.1. However for  $D(f) \neq \mathbf{R}^n$  some new situations arise. Above all it is obvious that  $\partial f(x)$  does not exist at any point  $x \notin D(f)$  (inequality (9) does not hold for any  $a$  for  $x + y \in D(f)$ ). At boundary points of  $D(f)$  the subgradient may or may not exist (example (5) and examples (6)-(8), respectively). But for the interior points of  $D(f)$  we have a result similar to Lemmas 6 and 8 of Section 5.1.

**LEMMA 3.** For  $x \in D(f)^0$  the set  $\partial f(x)$  is nonempty, convex, closed and bounded, and for any bounded set in  $D(f)^0$  the subgradients are uniformly bounded.  $\square$

The basic lemmas concerning the rules for calculating subgradients in Section 5.1 still hold under certain additional assumptions.

**LEMMA 4** (Moreau-Rockafellar). Let  $f_1(x), f_2(x)$  be convex functions, let  $f(x) = f_1(x) + f_2(x)$  and

$$D(f_1)^0 \cap D(f_2)^0 \neq \emptyset. \quad (10)$$

Then

$$\partial f(x) = \partial f_1(x) + \partial f_2(x). \quad \square \quad (11)$$

By induction, we get an extension of (11) to the case of  $m$  functions: if  $f_1(x), \dots, f_m(x)$  are convex and

$$D(f_1)^0 \cap \cdots \cap D(f_{m-1})^0 \cap D(f_m)^0 \neq \emptyset, \quad (12)$$

then

$$\partial(f_1(x) + \cdots + f_m(x)) = \partial f_1(x) + \cdots + \partial f_m(x). \quad (13)$$

The Moreau-Rockafellar lemma is a powerful tool for proving various results related to convex analysis and the theory of extremum problems. For the time being we limit ourselves to an important example, namely, recall that a set  $K \subset \mathbf{R}^n$  is called a *cone* if  $x \in K$  implies  $\lambda x \in K$  for any  $\lambda > 0$ . Examples of cones are a subspace, a halfspace, a right circular cone  $K = \{x \in \mathbf{R}^n : x_n \geq \sqrt{x_1^2 + \cdots + x_{n-1}^2}\}$  and a polyhedral cone  $K = \{x \in \mathbf{R}^n : (a^i, x) \leq 0, i = 1, \dots, m\}$  and in particular, the nonnegative orthant  $K = \{x : x \geq 0\}$  (Fig. 36). We have dealt with cones earlier: e.g., the set of sup-

porting vectors to a convex set (see Section 7.1) is a cone (called the *supporting cone*). For any set  $Q \subset \mathbf{R}^n$  one can consider the cone generated by it,  $K = \{x: x = \lambda y, \lambda > 0, y \in Q\}$ . Such are, for instance, the cone generated by the feasible directions, or the tangent cone (see Sections 7.1 and 8.1).

The cone  $K^*$  is said to be *conjugate* to the cone  $K \subset \mathbf{R}^n$  if (Fig. 36)

$$K^* = \{a \in \mathbf{R}^n : (a, x) \geq 0 \ \forall x \in K\}. \quad (14)$$

The cone  $-K^*$  is sometimes called the *polar* of the cone  $K$ .

**LEMMA 5** (Dubovitskij-Milyutin). Let  $K_1, \dots, K_m$  be convex cones in  $\mathbf{R}^n$ ,  $K = K_1 \cap \dots \cap K_m$  and

$$K_1^0 \cap \dots \cap K_{m-1}^0 \cap K_m^0 \neq \emptyset. \quad (15)$$

Then

$$K^* = K_1^* + \dots + K_m^*. \quad (16)$$

**PROOF.** Introduce  $f_i(x) = \delta_{K_i}(x)$ ,  $f(x) = \delta_K(x)$ , where  $\delta_Q(x)$  is the indicator function of the set  $Q$ :

$$\delta_Q(x) = \begin{cases} 0, & x \in Q, \\ \infty, & x \notin Q \end{cases} \quad (17)$$

(see Exercise 5). Then  $f(x) = f_1(x) + \dots + f_m(x)$ ,  $\partial f_i(0) = K_i^*$ ,  $\partial f(0) = K^*$  (see Exercise 6),  $D(f_i) = K_i$ ,  $D(f) = K$ . Hence (12) follows from (15) and (16) follows from (13).  $\square$ .

Condition (15) cannot generally be dropped (Exercise 4). However, in a very crucial special case (when all the  $K_i$  are closed halfspaces) one can get by without it.

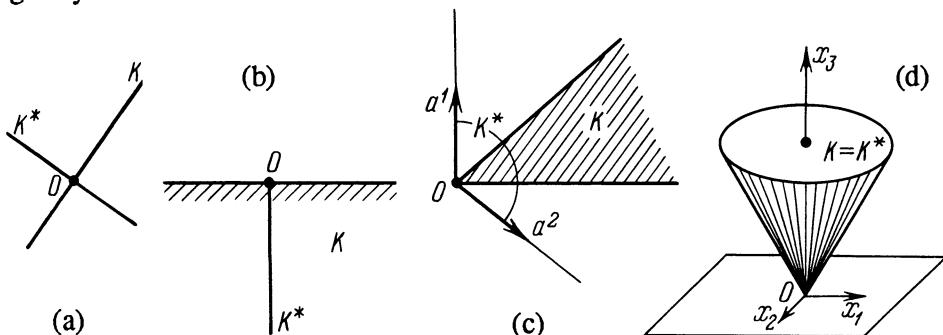


Fig. 36 Cones and conjugate cones: (a) subspace; (b) half-space; (c) polyhedral cone; (d) right circular cone.

**LEMMA 6 (Farkas).** Let  $K_i = \{x \in \mathbf{R}^n : (a^i, x) \geq 0\}$ ,  $i = 1, \dots, m$ ,  $K = K_1 \cap \dots \cap K_m \neq \emptyset$ . Then

$$K^* = K_1^* + \dots + K_m^* = \left\{ \sum_{i=1}^m y_i a^i, y_i \geq 0 \right\}.$$

Let  $A$  be an  $m \times n$  matrix whose rows are  $a^i$ . Then the required result can be written in the following form:

$$\begin{aligned} \text{if } K &= \{x \in \mathbf{R}^n : Ax \geq 0\}, \\ \text{then } K^* &= \{x \in \mathbf{R}^n : x = A^T y, y \geq 0, y \in \mathbf{R}^m\}. \end{aligned} \quad (18)$$

Using the same terms, we can write Lemma 1 of Section 8.1 in the following form:

$$\begin{aligned} \text{if } K &= \{x \in \mathbf{R}^n : Ax = 0\}, \\ \text{then } K^* &= \{x \in \mathbf{R}^n : x = A^T y, y \in \mathbf{R}^m\}. \end{aligned} \quad (19)$$

We can prove (18) following the same scheme as in proving (19) (see the proof of Lemma 1 of Section 8.1). Here, however, it is necessary to use the fact that a polyhedral cone of the form  $\{x = A^T y, y \geq 0\}$  is closed, and this fact requires a special proof.  $\square$

Farkas' lemma implies that if all the cones  $K_i$  are polyhedral, then in the Dubovitskij-Milyutin lemma one can drop condition (15). Similarly, one can strengthen the Moreau-Rockafellar lemma. Farkas' lemma makes it possible to write out the form of supporting vectors to a polyhedral set.

**LEMMA 7.** Let  $Q = \{x \in \mathbf{R}^n, (a^i, x) \leq b_i, i = 1, \dots, m\}$ ,  $x^* \in Q$ , and let  $I^* = \{i : (a^i, x^*) = b_i\}$  be the set of active constraints at  $x^*$ . Furthermore, let  $K = \{c : (c, x - x^*) \geq 0 \text{ for all } x \in Q\}$  be the cone of supporting vectors to  $Q$  at  $x^*$ . Then

$$K = \left\{ \sum_{i \in I^*} y_i a^i, y_i \geq 0, i \in I^* \right\}. \quad (20)$$

**PROOF.** Let  $\Gamma = \{z : z = \lambda(x - x^*), \lambda \geq 0, x \in Q\}$  be the cone generated by all feasible directions (see Section 7.1). Then  $\Gamma = \{z : (a^i, z) \leq 0, i \in I^*\}$ . By Farkas' lemma,  $\Gamma^* = \{\sum_{i \in I^*} y_i a^i, y_i \geq 0\}$ . If  $c \in \Gamma^*$ , then by the definition of the conjugate cone we have  $\lambda(c, x - x^*) \geq 0$  for all  $x \in Q$ ,  $\lambda \geq 0$ , which is equivalent to  $(c, x - x^*) \geq 0$  for all  $x \in Q$ , i.e.,  $\Gamma^* \subset K$ . The reverse inclusion is obvious.  $\square$

Finally, let us write the expression for the supporting cone for a set defined by a convex function.

**LEMMA 8.** Let  $f(x)$  be a convex function, and let  $x^* \in D(f)^0$ ,  $Q = \{x: f(x) \leq 0\}$  and  $\inf f(x) < 0$ . Then  $Q^* = \{c: (c, x - x^*) \geq 0 \text{ for all } x \in Q\} = \{\lambda \partial f(x^*), \lambda \geq 0\}$  if  $f(x^*) = 0$ , and  $Q^* = \{0\}$  if  $f(x^*) < 0$ .

The proof follows directly from the definition of the subgradient and Lemmas 2 and 6 (for  $m = 1$ ).  $\square$

### Exercises

1. Prove that if  $Q_\alpha = \{x: f(x) \leq \alpha\}$  is nonempty and bounded for some  $\alpha \in \mathbf{R}^1$  and for convex  $f(x)$  with  $D(f)^0 \supset Q_\alpha$ , then  $Q_\alpha$  is bounded for all  $\alpha < \infty$  (cf. Lemma 1 in Section 5.2). *Hint:* Use Lemma 2.
2. Prove Lemma 4. *Hint:* The inclusion  $\partial f(x) \supset \partial f_1(x) + \partial f_2(x)$  is obvious; to prove the reverse inclusion, consider the two sets in  $\mathbf{R}^{n+1}$ :

$$Q_1 = \{\alpha \in \mathbf{R}^1, z \in \mathbf{R}^n : \alpha \geq f_1(x+z) - f_1(x)\},$$

$$Q_2 = \{\alpha \in \mathbf{R}^1, z \in \mathbf{R}^n : \alpha < (\partial f(x), z) - f_2(x+z) + f_2(x)\}$$

and invoke the separation theorem.

3. Prove that an orthant and a right circular cone are self-conjugate, i.e.,  $K = K^*$ .

4. Let

$$K_1 = \{x \in \mathbf{R}^n : (a, x) > 0\} \cup \{0\},$$

$$K_2 = \{x : (a, x) = 0\}, \quad a \neq 0, \quad a \in \mathbf{R}^n.$$

Prove that in this case the equality  $(K_1 \cap K_2)^* = K_1^* + K_2^*$  is false.

5. Show that the indicator function of a convex set  $Q$  (17) is convex and has a subgradient at any point in  $Q$ , and  $\partial \delta_Q(x)$  coincides with the cone of vectors supporting to  $Q$  at  $x$ .

6. Let  $K$  be a convex cone. Show that  $\partial \delta_K(0) = K^*$ .

### 9.1.2 The Kuhn-Tucker Theorem

To begin, we give an unconstrained minimum criterion for a convex function, a particular case of which is Theorem 1 of Section 5.2.

**LEMMA 9.** Let  $f(x)$  be a convex function. Then  $x^*$  is a global minimum point of  $f(x)$  on  $\mathbf{R}^n$  iff

$$0 \in \partial f(x^*). \tag{21}$$

The proof follows immediately from the definition of the subgradient. Note that (21) implies the existence of the subgradient at a minimum point.  $\square$

**LEMMA 10.** Let  $f(x)$  be a convex function,  $\varepsilon > 0$ ,  $x^* \in D(f)$  and  $f(x) \geq f(x^*)$  for all  $x$  such that  $\|x - x^*\| \leq \varepsilon$ . Then  $f(x) \geq f(x^*)$  for all  $x$ .

**PROOF.** Let

$$\|x - x^*\| > \varepsilon, \quad \lambda = \varepsilon/\|x - x^*\| < 1, \quad x_\varepsilon = \lambda x + (1 - \lambda)x^*.$$

Then  $\|x_\varepsilon - x^*\| = \varepsilon$ , so that  $f(x_\varepsilon) \geq f(x^*)$ ; but from (1) we have that

$$f(x_\varepsilon) \leq \lambda f(x) + (1 - \lambda)f(x^*) ,$$

i.e.,  $f(x) \geq \lambda^{-1}(f(x_\varepsilon) - f(x^*)) + f(x^*) \geq f(x^*)$ .  $\square$

The assumption  $x^* \in D(f)$  is essential (otherwise, the expression  $f(x_\varepsilon) - f(x^*)$  in the foregoing calculation can be equal to  $\infty - \infty$  and becomes meaningless). For example, for the function (5) the point  $x = 2$  is not a global minimum, although in a neighborhood of it,  $f(x) = f(2) = \infty$ . Here caution should be exercised in making the computations involving expressions which may go into infinity.

Now let us examine a general problem of convex programming of the form

$$\begin{aligned} \min f(x), \quad x &\in \mathbf{R}^n, \\ g_i(x) \leq 0, \quad i &= 1, \dots, m, \\ x &\in Q. \end{aligned} \tag{22}$$

Furthermore, we formulate necessary and sufficient extremum conditions for this problem.

**THEOREM 1** (Kuhn-Tucker). Let  $f(x)$ ,  $g_i(x)$ ,  $i = 1, \dots, m$ , be convex functions, let  $Q$  be a convex set,  $Q \subset D(f)^0$ ,  $Q \subset D(g_i)^0$ ,  $i = 1, \dots, m$ , and let Slater's condition be satisfied: we can find an  $x^0 \in Q$  such that

$$g_i(x^0) < 0, \quad i = 1, \dots, m. \tag{23}$$

Then the admissible point  $x^*$  is a global solution of (22) if and only if we can find  $y_i^* \geq 0$ ,  $i = 1, \dots, m$ , such that

$$\begin{aligned} y_i^* g_i(x^*) = 0, \quad i &= 1, \dots, m \quad \text{and} \quad L(x, y^*) \geq L(x^*, y^*) \\ &\forall x \in Q, \end{aligned} \tag{24}$$

where

$$y^* = (y_1^*, \dots, y_m^*) \in \mathbf{R}^m, \quad g(x) = (g_1(x), \dots, g_m(x))$$

and

$$L(x, y) = f(x) + (y, g(x)). \quad (25)$$

As well in the case of problems with equality constraints (see Section 8.1), we say that the function  $L(x, y)$  is the Lagrangian, the vectors  $y^*$  are the Lagrange multipliers,  $x$  are the primal variables,  $y$  are the dual variables, the condition  $y_i^* g_i(x^*) = 0$ ,  $i = 1, \dots, m$ , is the complementarity condition, the set of indices  $I^* = \{i : g_i(x^*) = 0\}$  is the set of active constraints. Obviously,  $y_i^* = 0$  for  $i \notin I^*$ . If the conditions of Theorem 1 hold for problem (22), we then apply such terms as “regular minimum point” or a “regular problem.” The Kuhn-Tucker theorem claims that for a regular minimum point  $x^*$  there are nonnegative Lagrange multipliers  $y^*$  satisfying the complementarity condition such that the Lagrange function attains a minimum on  $Q$  at  $x^*$  for  $y = y^*$ . Thus there arises the possibility of reducing a problem with the inequalities  $g_i(x) \leq 0$  to a problem of minimization without these constraints.

**PROOF. Sufficiency.** Let  $x$  be an arbitrary admissible point and let (24) be satisfied. Then

$$\begin{aligned} f(x) &\geq f(x) + (y^*, g(x)) = L(x, y^*) \geq L(x^*, y^*) \\ &= f(x^*) + (y^*, g(x^*)) = f(x^*) , \end{aligned}$$

i.e.,  $x^*$  is the global minimum point in (22). Note that Slater's condition (23) has not been used.

**Necessity.** Introduce the functions

$$\begin{aligned} f_0(x) &= \delta_Q(x) , \quad f_i(x) = \delta_{Q_i}(x) , \\ Q_i &= \{x : g_i(x) \leq 0\} , \quad i = 1, \dots, m , \\ F(x) &= f(x) + \sum_{i=0}^m f_i(x) , \end{aligned} \quad (26)$$

where  $\delta_Q(x)$  is the indicator function (17). Then  $F(x) = f(x)$  if  $x$  is an admissible point, and  $F(x) = \infty$  otherwise. Therefore problem (22) is equivalent to unconstrained minimization of  $F(x)$ . The function  $F(x)$  is convex, and hence by Lemma 9,  $0 \in \partial F(x^*)$ . The point  $x^0$  in (23) is such that  $x^0 \in D(f)^0$ ,  $x^0 \in D(g_i)^0$ ,  $x^0 \in Q_i^0$  by Lemma 2, and therefore  $x^0 \in D(f_i)^0$ ,  $i = 1, \dots, m$ . Hence

$$D(f)^0 \cap D(f_0) \cap D(f_1)^0 \cap \cdots \cap D(f_m)^0 \neq \emptyset ,$$

V(6)

and the Moreau-Rockafellar lemma is applicable to  $F(x)$ :

$$\partial F(x^*) = \partial f(x^*) + \partial f_0(x^*) + \cdots + \partial f_m(x^*).$$

But  $\partial f_0(x^*) = \partial \delta_Q(x^*)$  is the cone of supporting vectors to  $Q$  at  $x^*$  (see Exercise 5),

$$\partial f_i(x^*) = \{y_i \partial g_i(x^*), y_i \geq 0\}, \quad i \in I^*;$$

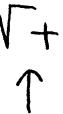
$$\partial f_i(x^*) = 0, \quad i \notin I^*$$

(Lemma 8). Thus we can find  $y_i^* \geq 0$ ,  $i \in I^*$ , such that

$$0 \in \partial f(x^*) + \partial \delta_Q(x^*) + \sum_{i \in I^*} y_i^* \partial g_i(x^*). \quad (27)$$

Introduce the vector  $y^* = (y_1^*, \dots, y_m^*)$ , where the  $y_i^*$  have been defined for  $i \in I^*$ ,  $y_i^* = 0$  for  $i \notin I^*$ , and the function

$$\phi(x) = L(x, y^*) + \delta_Q(x) = f(x) + \delta_Q(x) \sum_{i \in I^*} y_i^* \partial g_i(x^*). \quad (28)$$



The Moreau-Rockafellar lemma can be applied again since

$$x^* \in D(f)^0, \quad x^* \in D(g_i)^0, \quad i \in I^*, \quad x^* \in D(\delta_Q),$$

and we have that

$$\partial \phi(x^*) = \partial f(x^*) + \partial \delta_Q(x^*) + \sum_{i \in I^*} y_i^* \partial g_i(x^*).$$

Thus, (27) has the form  $0 \in \partial \phi(x^*)$ , and by Lemma 9  $x^*$  is the unconstrained minimum point of  $\phi(x)$ . This is equivalent (see (28)) to  $x^*$  being a minimum point of  $L(x, y^*)$  on  $Q$ .  $\square$

Note that from (27) we obtain an extremum condition in the subgradient form:

$$\begin{aligned} (\partial_x L(x^*, y^*), x - x^*) &\geq 0, \quad \forall x \in Q, \\ \partial_x L(x^*, y^*) &= \partial f(x^*) + \sum_{i \in I^*} y_i^* \partial g_i(x^*), \end{aligned} \quad (29)$$

which is a generalization of Theorem 3 in Section 7.1, in which the problem without inequality constraints was examined. But if  $Q = \mathbf{R}^n$ , then (29) becomes

$$0 \in \partial_x L(x^*, y^*). \quad (30)$$

Many other proofs for the Kuhn-Tucker theorem are known (e.g., those based directly on the separation theorem). The simple proof we have given demonstrates convincingly the effectiveness of the convex analysis technique. If one tries to use the extremum conditions from Chapter 7 and employ the set defined by all the constraints, one then need to write the support vector to this set. This is precisely what constitutes the nontrivial part of the Kuhn-Tucker theorem.

Slater's condition plays the same role as the regularity condition does in the problem with equality constraints (see Section 8.1). If this condition is not satisfied, the admissible set may turn out to be too "meager" and the Kuhn-Tucker theorem will be false. For instance, if problem (22) has the form (see Fig. 34)

$$\begin{aligned} \min x_2, \quad x \in \mathbf{R}^2, \\ (x_1 - 1)^2 + x_2^2 - 1 \leq 0, \\ (x_1 + 1)^2 + x_2^2 - 1 \leq 0, \end{aligned} \tag{31}$$

then  $x^* = 0$ ,  $I^* = \{1, 2\}$  and, as is not hard to verify, there exist no  $y_1^*$ ,  $y_2^*$  such that (30) is satisfied.

If the problem contains linear constraints, it is convenient to relate them to a set  $Q$ . In particular, if there are no other constraints, i.e., the problem has the form

$$\begin{aligned} \min f(x), \quad x \in \mathbf{R}^n, \\ (a^i, x) \leq b_i, \quad i = 1, \dots, m, \end{aligned} \tag{32}$$

it is preferable to assume that all these constraints define the polyhedral set  $Q$  and, in addition, to use Lemma 7 concerning the form of support vectors for such a set.

**THEOREM 2.** If  $f(x)$  is convex and  $D(f)^0$  contains the admissible set, then a necessary and sufficient extremum condition in (32) has the form: there are  $y_i^* \geq 0$ ,  $i \in I^* = \{i: (a^i, x^*) = b_i\}$  such that

$$\sum_{i \in I^*} y_i^* a^i \in -\partial f(x^*). \quad \square \tag{33}$$

We emphasize the fact that in this case Slater's condition is not required, i.e., the admissible set in (32) may or may not have an interior point.

The Kuhn-Tucker theorem is often written in a somewhat different form, viz. in terms of a saddle point. To do this, we introduce the appropriate notions. Let  $Q \subset \mathbf{R}^n$ ,  $S \subset \mathbf{R}^m$  be two sets,  $\phi: Q \times S \rightarrow \mathbf{R}^1$ . We call a pair  $x^* \in Q$ ,  $y^* \in S$  a saddle point for the function  $\phi(x, y)$  on  $Q \times S$  if

$$\phi(x^*, y) \leq \phi(x^*, y^*) \leq \phi(x, y^*) \quad \forall x \in Q \text{ and } \forall y \in S. \tag{34}$$

In other words,  $x^*$  is a minimum point of  $\phi(x, y^*)$  in  $x$  on  $Q$ , and  $y^*$  is a maximum point of  $\phi(x^*, y)$  in  $y$  on  $S$ . If the expressions written below are defined, the equality

$$\min_{x \in Q} \max_{y \in S} \phi(x, y) = \max_{y \in S} \min_{x \in Q} \phi(x, y) = \phi(x^*, y^*) \quad (35)$$

is equivalent to (34), i.e., the existence of a saddle point implies that the operations of minimization and maximization can be interchanged.

**THEOREM 3** (Kuhn-Tucker). Under the conditions of Theorem 1  $x^*$  is a solution of problem (22) if and only if the pair  $x^*, y^*$  for some  $y^* \geq 0$  is a saddle point of  $L(x, y)$  on  $Q \times \mathbf{R}_+^m$ , i.e.,

$$L(x^*, y) \leq L(x^*, y^*) \leq L(x, y^*) \quad \forall x \in Q \text{ and } \forall y \geq 0. \quad (36)$$

**PROOF.** Let  $x^*$  be a solution of (22). By Theorem 1, we can find a  $y^* \geq 0$  such that  $(y^*, g(x^*)) = 0$  and  $L(x, y^*) \geq L(x^*, y^*)$  for all  $x \in Q$ . But then  $L(x^*, y^*) \geq f(x^*) + (y, g(x^*)) = L(x^*, y)$  for any  $y \geq 0$  since  $g(x^*) \leq 0$ . Thus the pair  $x^*, y^*$ , where  $x^*$  is a solution of (22) and  $y^*$  are the Lagrange multipliers, is a saddle point of  $L(x, y)$  on  $Q \times \mathbf{R}_+^m$ .

Conversely, let  $x^*, y^*$  be a saddle point. Then  $L(x^*, y^*) \geq L(x^*, y)$  implies that  $(y, g(x^*)) \leq (y^*, g(x^*))$  for all  $y \geq 0$ . This is possible only if  $g(x^*) \leq 0$ ,  $(y^*, g(x^*)) = 0$ . Hence for any admissible  $x$  one has

$$L(x^*, y^*) = f(x^*) \leq L(x, y^*) = f(x) + (y^*, g(x)) \leq f(x),$$

which means that  $x^*$  is a solution of problem (22).  $\square$

## Exercises

7. Derive conditions (8) in Section 7.1 from Theorem 2. *Hint:* Write the constraints  $(a^i, x) = b_i$  in the form  $(a^i, x) \leq b_i$ ,  $-(a^i, x) \leq b_i$ .

8. Check that both functions  $\phi(x, y) = xy$  and  $\phi(x, y) = -xy$ ,  $x \in \mathbf{R}^1$ ,  $y \in \mathbf{R}^1$ , have a unique point  $\{0, 0\}$  on  $\mathbf{R}^1 \times \mathbf{R}^1$ .

9. Use the example in Exercise 8 to see that (34) does not imply the equalities

$$X^* = \operatorname{Argmin}_{x \in Q} \phi(x, y^*), \quad Y^* = \operatorname{Argmax}_{y \in S} \phi(x^*, y),$$

where  $X^* \times Y^*$  is the set of saddle points.

### 9.1.3. Duality

In the formulation of Theorem 3 the primal and dual variables are symmetric. Hence one might expect that a similar symmetry exists for optimization problems too, i.e., that (36) is an extremum condition not only for the initial problem (22) but also for an optimization problem with respect to dual variables. Such a problem may be obtained from the following considerations. Let

$$\phi(x) = \sup_{y \geq 0} L(x, y), \quad (37)$$

Then it is obvious that

$$\phi(x) = \begin{cases} f(x) & \text{if } g_i(x) \leq 0, i = 1, \dots, m, \\ \infty, & \text{otherwise.} \end{cases}$$

Hence the initial problem can be written as

$$\min_{x \in Q} \phi(x). \quad (38)$$

We proceed analogously and interchange the roles of the variables and the operations of minimization and maximization: introduce

$$\psi(y) = \inf_{x \in Q} L(x, y) \quad (39)$$

(possibly  $\psi(y) = -\infty$  for some  $y$ ) and consider the problem

$$\max_{y \geq 0} \psi(y). \quad (40)$$

Problem (40) is called the dual problem, and (38) or (22) is called the primal problem.

**THEOREM 4** (duality theorem). The following duality relations hold:

(a) For any admissible  $x$  and  $y$  (i.e., for  $x \in Q$ ,  $g_i(x) \leq 0$ ,  $i = 1, \dots, m$ ,  $y \geq 0$ ), one has

$$f(x) \geq \psi(y). \quad (41)$$

(b) If the primal problem is regular,  $x^*$  is a solution to the problem, and  $y^*$  are the Lagrange multipliers, then  $y^*$  is a solution of (40) and

$$f(x^*) = \psi(y^*). \quad (42)$$

(c) If (42) holds for admissible  $x^*$ ,  $y^*$ , then  $x^*$  is a solution of the primal problem and  $y^*$  is a solution of the dual problem.

**PROOF.** (a) if  $x \in Q$ ,  $g(x) \leq 0$ ,  $y \geq 0$ , then

$$f(x) \geq f(x) + (y, g(x)) = L(x, y) \geq \inf_{x' \in Q} L(x', y) = \psi(y).$$

(b) Let  $x^*$  be a solution of (22) and let  $y^*$  be the Lagrange multipliers. Then by Theorem 3 one has

$$\overline{\psi}(y^*) = \inf_{x \in Q} L(x, y^*) = L(x^*, y^*) \geq L(x^*, y) \geq \inf_{x \in Q} L(x, y) = \overline{\psi}(y)$$

for all  $y \geq 0$ , i.e.,  $y^*$  is a solution of (40) and, furthermore, since  $L(x^*, y^*) = f(x^*)$ , then  $\psi(y^*) = f(x^*)$ .

(c) Let  $g(x^*) \leq 0$ ,  $x^* \in Q$ ,  $y^* \geq 0$  and  $f(x^*) = \psi(y^*)$ . Then for arbitrary admissible  $x$ ,  $y$ , by (41),  $f(x) \geq \psi(y^*) = f(x^*) \geq \psi(y)$ , i.e.,  $x^*$ ,  $y^*$  are solutions of (22) and (40).  $\square$

Let us consider now two examples. For the problem in  $\mathbf{R}^2$

$$\min x_1, x_2 \leq 0 \quad (43)$$

we have

$$L(x, y) = x_1 + x_2 y, \quad \psi(y) = \inf_{x \in \mathbf{R}^2} L(x, y) \equiv -\infty.$$

In this case neither the primal problem nor the dual problem has a solution. Let

$$f(x) = 1/x, \quad x \in \mathbf{R}^1, \quad g(x) = -x \leq 0. \quad (44)$$

Then  $L(x, y) = 1/x - xy$ ,  $\psi(y) = -\infty$  for  $y > 0$ ,  $\psi(0) = 0$ . Here the dual problem has a solution  $y^* = 0$  but the primal problem has none. Finally, if the problem is

$$\min x, \quad x^2 \leq 0, \quad x \in \mathbf{R}^1, \quad (45)$$

then  $L(x, y) = x + yx^2$ ,  $\psi(y) = -(4y)^{-1}$ , and the primal problem has the solution  $x^* = 0$  but the dual problem has none. All these "pathological" examples show that in the general case the relationship of the primal problem to the dual problem can be arbitrary. However, in the regular case, by Theorem 4(b), both problems have solutions simultaneously, and their optimal values are equal.

The following factors make the duality theorem more useful than usual extremum conditions—such as Theorems 1 to 3: (1) it is possible to reduce the initial problem to a simpler problem. Thus, if  $m \ll n$ , then the dimension of the dual problem (equal to  $m$ ) is substantially lower than that of the primal problem; (2) inequality (41) makes it possible to obtain a lower bound for the minimum in (22) and estimate thereby the accuracy of the

approximate solution. Of course, how fruitful the dual approach is depends to a great extent on how easy the computation of  $\psi(y)$  is. In a number of cases (in particular, for problems of linear, quadratic, separable and geometric programming, see Chapters 10 and 11) the dual approach proves to be very efficient, indeed.

Note that we shall sometimes write the dual problem (40) in a different form. First, the maximum of  $\psi(y)$  can be attained only at points such that  $\psi(y) \neq -\infty$ ; therefore (40) is equivalent to the problem

$$\max \phi(y), \quad y \geq 0, \quad y \in D(\psi), \quad (46)$$

where  $D(\psi) = \{y: \psi(y) > -\infty\}$ . Second, it is more customary to deal with minimization problems than with maximization problems. If we introduce

$$\theta(y) = -\psi(y), \quad (47)$$

then instead of (40) we obtain the convex (Exercise 10) problem

$$\min \theta(y), \quad y \geq 0, \quad (48)$$

or instead of (46) the problem

$$\min \theta(y), \quad y \geq 0, \quad y \in D(\theta); \quad (49)$$

thus the duality relation (41) takes on the form

$$f(x) + \theta(y) \geq 0. \quad (50)$$

To conclude, we note that we obtain differing dual problems depending on whether the constraints are written in the form  $g_i(x) \leq 0$  or related to the set  $Q$ . In general, for each optimization problem there exist many dual problems. The general theory is not, however, of our concern at the moment.

## Exercises

**10.** Prove the following properties of  $\psi(y)$ :

- (a) the set  $D(\psi) = \{y: \psi(y) > -\infty\}$  is convex, the function  $\psi(y)$  is concave on  $D(\psi)$ ;
- (b) if  $f(x)$ ,  $g_i(x)$  are continuous,  $Q$  is closed and bounded, then  $D(\psi) = \mathbf{R}^m$  and  $\psi(y)$  is continuous;
- (c) if  $f(x)$ ,  $g_i(x)$  are convex,  $Q$  is convex,  $Q \subset D(f)^0$ ,  $Q \subset D(g_i)^0$ ,  $f(x)/\|x\| \rightarrow \infty$  for  $x \in Q$ ,  $\|x\| \rightarrow \infty$ , then  $\mathbf{R}_+^m \subset D(\psi)$ ;
- (d) if  $f(x)$  is strictly convex,  $g_i(x)$  are convex,  $Q$  is convex, closed and bounded, then  $\psi(y)$  is differentiable for  $y \geq 0$  and  $\nabla \psi(y) = g(x(y))$ , where  $x(y) = \arg \min_{x \in Q} L(x, y)$ ;

(e) if, moreover,  $f(x)$  is strongly convex, then  $\nabla\psi(y)$  satisfies a Lipschitz condition.

11. Write the duals of the problems:

$$(a) \min (c, x), \|x\| \leq 1;$$

(b)  $\min (c, x), \|x\|^2 \leq 1$ . Check that although the initial problems are equivalent, different dual problems are obtained.

#### 9.1.4 Existence, Uniqueness and Stability of a Solution

Using Lemma 2 and the result of Exercise 1, we can formulate the following version of Weierstrass's theorem for problem (22).

**THEOREM 5.** Let  $f(x)$ ,  $g_i(x)$ ,  $i = 1, \dots, m$ , be convex functions and let  $Q$  be convex and closed. For  $S = \{x \in Q: g_i(x) \leq 0, i = 1, \dots, m\}$  let  $S \subset D(f)^0$ ,  $S \subset D(g_i)^0$ ,  $i = 1, \dots, m$ , and let the set  $\{x \in S: f(x) \leq \alpha\}$  be nonempty and bounded for some  $\alpha$ . Then a solution of (22) exists.  $\square$

As usual, a unique solution can be guaranteed for a strictly convex function  $f(x)$ . Moreover, if  $Q$  is strictly convex, and  $\|\partial_x L(x, y^*)\| \geq \varepsilon > 0$  for all  $x \in Q$  and also the conditions of Theorem 1 hold, then it is easy to prove that the solution is unique. Finally, if any of the  $g_i(x)$  is strictly convex and the corresponding Lagrange multiplier  $y_i^*$  is positive, then the solution is unique, too.

In regard to the dual problem, regularity assumptions do not suffice for the solution to be unique. For example, in the problem

$$\begin{aligned} \min x, \quad x \in \mathbf{R}^1, \\ g_1(x) = -x \leq 0, \quad g_2(x) = x^2 - 2x \leq 0 \end{aligned} \tag{51}$$

the Lagrange multipliers are determined nonuniquely. Roughly, this is because the first constraint is redundant (i.e., with or without it, the admissible set remains the same). At the same time, under regularity conditions one may assert that the solutions of the dual problem are bounded.

We now turn to a stability analysis. We need the following result concerning the continuous dependence of the set of solutions of a system of convex inequalities on the right sides.

**LEMMA 11.** Let  $g_i(x)$ ,  $i = 1, \dots, m$  be convex functions, let the set

$$S = \{x: g_i(x) \leq 0, i = 1, \dots, m\} \tag{52}$$

be convex and bounded, and let  $D(g_i)^0 \supset S$ . Then

(a) the set

$$S_\varepsilon = \{x: g_i(x) \leq \varepsilon_i, i = 1, \dots, m\} \tag{53}$$

is nonempty and bounded for any  $\varepsilon > 0$ ,  $\varepsilon = \{\varepsilon_1, \dots, \varepsilon_m\}$  and  $\Delta(S_\varepsilon, S) \rightarrow 0$  as  $\varepsilon \rightarrow +0$ ;

(b) if, in addition, there exists a point  $x^0$  such that  $g_i(x^0) \leq -\delta < 0$ , then  $S_\varepsilon$  is nonempty and bounded for any  $\varepsilon_i \geq g_i(x^0)$ ,  $i = 1, \dots, m$ , and

$$\rho(x_\varepsilon, S) \leq c \left( \max_{1 \leq i \leq m} \varepsilon_i \right)_+ \quad \forall x_\varepsilon \in S_\varepsilon, \quad c = \frac{1}{\delta} \max_{x \in S} \|x - x^0\|. \quad \square \quad (54)$$

Here  $\Delta(S_\varepsilon, S)$  is the Hausdorff distance between the sets  $S_\varepsilon$  and  $S$ , i.e.,

$$\Delta(S_\varepsilon, S) = \max \left\{ \max_{x \in S} \rho(x, S_\varepsilon), \max_{x_\varepsilon \in S_\varepsilon} \rho(x_\varepsilon, S) \right\}.$$

Lemma 11 makes it possible to obtain directly a result on weak stability (see Section 1.3) of a convex programming problem towards a perturbation of the constraints.

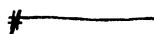
**THEOREM 6.** Let

$$X^* = \operatorname{Argmin}_{x \in S} f(x), \quad X_\varepsilon^* = \operatorname{Argmin}_{x \in S_\varepsilon} f(x),$$

where  $S$  and  $S_\varepsilon$  have the form (52) and (53),  $f(x)$  and  $g_i(x)$  are convex and continuous on  $D(f)^0$  and  $D(f)^0 \supset S$ ,  $D(g_i)^0 \supset S$ ,  $i = 1, \dots, m$ , and also  $X^*$  is nonempty and bounded. Then

(a) for all  $\varepsilon > 0$  the set  $X_\varepsilon^*$  is nonempty and bounded and  $\Delta(X_\varepsilon^*, X^*) \rightarrow 0$ ,  $f_\varepsilon^* \rightarrow f^*$  as  $\varepsilon \rightarrow +0$ , where  $f^* = f(x^*)$ ,  $x^* \in X^*$ ,  $f_\varepsilon^* = f(x_\varepsilon^*)$ ,  $x_\varepsilon^* \in X_\varepsilon^*$ ;

(b) if there exists  $x^0$ :  $g_i(x^0) < 0$ ,  $i = 1, \dots, m$ , then  $X_\varepsilon^*$  is nonempty and bounded for any  $\varepsilon_i \geq g_i(x^0)$ ,  $i = 1, \dots, m$ , and  $\Delta(X_\varepsilon^*, X^*) \rightarrow 0$ ,  $f_\varepsilon^* \rightarrow f^*$  as  $\varepsilon \rightarrow 0$ .

 **LEAD 12**

**PROOF.** The  $X^*$  can be written in the form  $X^* = \{x: g_i(x) \leq 0, i = 0, 1, \dots, m\}$ , where  $g_0(x) = f(x) - f^*$ . By Lemma 11, the set  $X_\varepsilon = \{x: g_i(x) \leq \varepsilon_i, i = 0, 1, \dots, m\}$  (where  $\varepsilon_0 = 0$  in case (a) and  $\varepsilon_0 = f(x^0) - f^*$  in case (b)) is bounded and nonempty. But minimization of  $f(x)$  on  $S_\varepsilon$  is equivalent to minimization of  $f(x)$  on  $X_\varepsilon$ . The set  $X_\varepsilon$ , as we have shown, is bounded and nonempty; by the convexity and continuity of  $f(x)$  and  $g_i(x)$  it is convex and closed. Hence the minimum of  $f(x)$  on  $X_\varepsilon$  obtains, i.e.,  $X_\varepsilon^*$  is nonempty and bounded. Since  $f(x)$  is continuous in a neighborhood of  $X^*$ , we obtain  $f_\varepsilon^* \rightarrow f^*$ . Writing  $X_\varepsilon^*$  in the form  $X_\varepsilon^* = \{x: g_i(x) \leq \varepsilon_i, i = 0, 1, \dots, m\}$ , where  $\varepsilon_0 = f_\varepsilon^* - f^*$ , and again applying Lemma 11, we obtain  $\Delta(X_\varepsilon^*, X^*) \rightarrow 0$ . 

We are not going to discuss now the strong stability of solutions, stability towards perturbations of the objective function, stability of the

Lagrange multipliers, more general forms of perturbations, or any other results concerning stability. We only note that the duality theorem is of great use in the investigation of such problems. For example, instead of problem (22) we are examining, say, the perturbed problem

$$\min f(x), \quad g_i(x) \leq \varepsilon_i, \quad i = 1, \dots, m, \quad x \in Q. \quad (55)$$

We form its Lagrange function

$$L_\varepsilon(x, y) = f(x) + (y, g(x) - \varepsilon) = L(x, y) - (\varepsilon, y) \quad (56)$$

and write out the dual problem

$$\max_{y \geq 0} \psi_\varepsilon(y), \quad \psi_\varepsilon(y) = \inf_{x \in Q} L_\varepsilon(x, y) = \psi(y) - (\varepsilon, y). \quad (57)$$

Thus the problem has been reduced to an investigation of the stability of the dual problem when the objective function is perturbed.

### Exercise

12. Consider the example in  $\mathbf{R}^2$ , where the initial problem:  $\min x_1, -x_1 \leq 0$ , is perturbed:  $\min (x_1 + \varepsilon_1 x_2), -x_1 - \varepsilon_2 x_2 \leq 0$ . Check that in the initial problem a solution exists ( $x_1^* = 0, x_2^*$  is arbitrary), but in the perturbed problem there is no solution for  $\varepsilon_1 \neq \varepsilon_2$ .

## 9.2 NONLINEAR PROGRAMMING (THEORY)

We shall consider the general problem of nonlinear programming

$$\begin{aligned} & \min f(x) \\ & g_i(x) \leq 0, \quad i = 1, \dots, r, \\ & g_i(x) = 0, \quad i = r+1, \dots, m, \end{aligned} \quad (1)$$

where the functions are assumed to be differentiable but not necessarily convex.

### 9.2.1 Necessary Conditions for a Minimum

For any admissible point  $x^*$  we introduce the sets of indices:

$$\begin{aligned} I^* &= \{i: g_i(x^*) = 0, i = 1, \dots, r\}, \\ I &= \{i: g_i(x^*) = 0, i = 1, \dots, m\}, \end{aligned} \quad (2)$$

characterizing the active indices ( $I^*$  relates to inequalities and  $I$  to all the constraints).

**THEOREM 1** (Karush-John). Let  $x^*$  be a local minimum point in (1) and let the functions  $f(x)$ ,  $g_i(x)$ ,  $i = 1, \dots, m$ , be continuous and differentiable in a neighborhood of  $x^*$ . There we can find  $y_0^*, y_i^*$ ,  $i \in I$ , not all equal to 0, such that  $y_0^* \geq 0$ ,  $y_i^* \geq 0$ ,  $i \in I^*$ , and

$$y_0^* \nabla f(x^*) + \sum_{i \in I} y_i^* \nabla g_i(x^*) = 0. \quad (3)$$

**PROOF.** One can assume that  $\nabla g_i(x^*)$ ,  $i = r+1, \dots, m$ , are linearly independent, because otherwise the assertion of the theorem is trivial (one can take  $y_0^* = 0$ ,  $y_i^* = 0$ ,  $i \in I^*$ ). We construct two sets in  $\mathbf{R}^{m+1}$  and denote the components of  $z \in \mathbf{R}^{m+1}$  by  $\{z_0, z_1, \dots, z_m\}$ :

$$A = \{z \in \mathbf{R}^{m+1} : z_0 = (\nabla f(x^*), s), z_i = (\nabla g_i(x^*), s), i \in I, s \in \mathbf{R}^n\},$$

$$B = \{z \in \mathbf{R}^{m+1} : z_0 < 0, z_i < 0, i \in I^*; z_i = 0, i = r+1, \dots, m\}.$$

Let us show that these sets are disjoint. Let the opposite be true. Then we can find an  $s \in \mathbf{R}^n$  such that

$$\begin{aligned} (\nabla f(x^*), s) &< 0, \\ (\nabla g_i(x^*), s) &< 0, \quad i \in I^*, \\ (\nabla g_i(x^*), s) &= 0, \quad i = r+1, \dots, m. \end{aligned} \quad (4)$$

By Lyusternik's theorem (Theorem 3 of Section 8.1) there are points  $x_\lambda$  such that

$$g_i(x_\lambda) = 0, \quad i = r+1, \dots, m, \quad x_\lambda = x^* + \lambda s + o(\lambda).$$

Then by (4) we have

$$f(x_\lambda) = f(x^*) + \lambda(\nabla f(x^*), s) + o(\lambda) < f(x^*)$$

for sufficiently small  $\lambda > 0$ . On the other hand, for the same reason,  $g_i(x_\lambda) < 0$  for  $i \in I^*$  and sufficiently small  $\lambda > 0$ . Finally,  $g_i(x_\lambda) < 0$  for  $i \notin I$  and small  $\lambda$  by the continuity of  $g_i(x)$  and the fact that  $g_i(x^*) < 0$  for  $i \in I$ . Thus, for small  $\lambda > 0$  the point  $x_\lambda$  satisfies all the constraints, and  $f(x_\lambda) < f(x^*)$ , which contradicts the definition of a local minimum. Hence,  $A$  and  $B$  are disjoint.

The sets  $A$  and  $B$  are obviously convex and nonempty. By the separation theorem there is a  $y^* \in \mathbf{R}^{m+1}$ ,  $y^* \neq 0$ , such that  $(y^*, z) \geq 0$ ,  $z \in A$ ,  $(y^*, z) \leq 0$ ,  $z \in B$  (one can take 0 on the right side of the inequalities

since  $A$  and  $B$  are cones). Denote the components of  $y^*$  by  $y_0^*, y_1^*, \dots, y_m^*$ . Then the inequality  $(y^*, z) \geq 0$  for  $z \in A$  implies that

$$y_0^*(\nabla f(x^*), s) + \sum_{i \in I} y_i^*(\nabla g_i(x^*), s) + \sum_{i \notin I} y_i^* z_i \geq 0$$

for all  $s \in \mathbb{R}^n$  and all  $z_i \in \mathbb{R}^1$ ,  $i \notin I$ . This is possible only if  $y_0^* \nabla f(x^*) + \sum_{i \in I} y_i^* \nabla g_i(x^*) = 0$  and  $y_i^* = 0$  for  $i \notin I$ .

Finally, since for all  $z \in B$

$$(y^*, z) = y_0^* z_0 + \sum_{i \in I^*} y_i^* z_i \leq 0$$

for any  $z_0 < 0$ ,  $z_i < 0$ ,  $i \in I^*$ , then  $y_0^* \geq 0$ ,  $y_i^* \geq 0$ ,  $i \in I^*$ .  $\square$

Another way of proving the theorem is suggested in Exercise 2.

We are now interested in the conditions (called regularity conditions) under which  $y_0^* \neq 0$  is guaranteed. Since the  $y_i^*$  are defined to within a positive multiple, then we may assume without loss of generality that  $y_0^* = 1$ . In other words, the necessary extremum condition takes on the form

$$L'_x(x^*, y^*) = 0, \quad y_i^* \geq 0, \quad i \in I^*, \quad (y^*, g(x^*)) = 0, \quad (5)$$

$$\begin{aligned} L(x, y) &= f(x) + (y, g(x)), \quad y^* = (y_0^*, \dots, y_m^*), \\ g(x) &= (g_1(x), \dots, g_m(x)). \end{aligned} \quad (6)$$

As earlier, we call the function  $L(x, y)$  the Lagrange function and the numbers  $y_i^*$ ,  $i = 1, \dots, m$ , the Lagrange multipliers.

The most elementary regularity condition is

**Regularity condition A:** the vectors  $\nabla g_i(x^*)$ ,  $i \in I$ , are linearly independent.

**THEOREM 2.** Let the conditions of Theorem 1 and regularity condition A be satisfied. Then we can find a  $y^* \in \mathbb{R}^m$  such that (5) is satisfied.

The proof follows directly from (3) since the assumption  $y_0^* = 0$  contradicts regularity condition A.  $\square$

Condition A is restrictive to a large extent and is not always satisfiable. **Regularity condition B:** the vectors  $\nabla g_i(x^*)$ ,  $i = r+1, \dots, m$ , are linearly independent and we can find a vector  $s^0 \in \mathbb{R}^n$  such that

$$(\nabla g_i(x^*), s^0) = 0, \quad i = r+1, \dots, m;$$

$$(\nabla g_i(x^*), s^0) < 0, \quad i \in I^*.$$

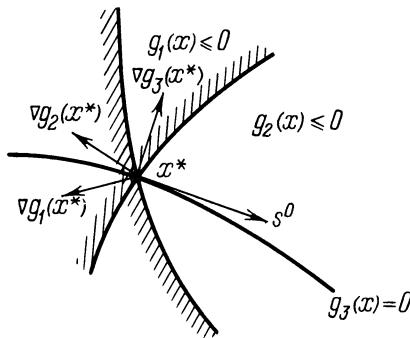


Fig. 37 Regularity condition B.

In other words, we can find an element in the tangent subspace to the equality constraints at the point  $x^*$  which leads strictly inside each of the sets  $g_i(x) \leq 0$ ,  $i \in I^*$  (Fig. 37). In particular, if the functions  $g_i(x)$ ,  $i = 1, \dots, m$ , are convex while constraints in the form of equalities are absent and Slater's condition (23) in Section 9.1 holds, then taking  $s^0 = x^0 - x^*$ , we find that condition B holds too. Slater's condition is convenient since it is not required to know the solution  $x^*$ .

**THEOREM 3.** When condition A is changed to condition B, the assertion of Theorem 2 remains valid.

**PROOF.** Suppose that  $y_0^* = 0$  and compute the scalar product of (3) by  $s^0$ . We obtain

$$\sum_{i \in I^*} y_i^* (\nabla g_i(x^*), s^0) = 0,$$

which together with the conditions  $(\nabla g_i(x^*), s^0) > 0$ ,  $y_i^* \geq 0$ ,  $i \in I^*$ , yield  $y_i^* = 0$ ,  $i \in I^*$ . Hence (3) becomes  $\sum_{i=r+1}^m y_i^* \nabla g_i(x^*) = 0$ ,  $y_i^*$  not all equal to 0, which contradicts the linear independence of the  $\nabla g_i(x^*)$ ,  $i = r+1, \dots, m$ .  $\square$

### Exercises

1. Compare the proof of Theorem 1 with the first proof of the rule of Lagrange multipliers (see Section 8.1). Which assertion was used there instead of the separation theorem?

2. Prove Theorem 1 following the scheme using penalty functions (cf. the third proof of Theorem 2 of Section 8.1). *Hint:* Introduce

$$f_k(x) = f(x) + \frac{1}{2}K \left[ \sum_{i=1}^r g_i(x)_+^2 + \sum_{i=r+1}^m g_i(x)^2 \right] + \|x - x^*\|^2.$$

Show that  $x^k$ , a minimum point of  $f_k(x)$  on  $Q = \{x: \|x - x^*\| \leq \varepsilon\}$ , lies inside  $Q$  for large  $K$  for large  $K$ , use the necessary minimum conditions without constraints, i.e.,  $\nabla f_k(x^k) = 0$ , and pass to the limit as  $K \rightarrow \infty$ . Show that if the minimum point  $x^*$  is locally unique, then the term  $\|x - x^*\|^2$  in  $f_k(x)$  can be discarded.

### 9.2.2 Sufficient Conditions for a Minimum

It is possible to assert in certain cases that  $x^*$  is a solution, employing only the first derivatives and not assuming that the problem is convex. This result has no analogue either in unconstrained minimization problems or in equality constraint problems, but it is close to the condition for a sharp minimum (see Section 7.1).

**THEOREM 4** (sufficient first-order conditions). Let  $x^*$  be an admissible point, let the functions  $f(x)$ ,  $g_i(x)$ ,  $i = 1, \dots, m$ , be differentiable at  $x^*$ , let  $n$  be the number of active constraints, and let their gradients (i.e.,  $\nabla g_i(x^*)$ ) be linearly independent. Suppose there are scalars  $y_i^*$ ,  $i \in I^*$  such that  $y_i^* > 0$ ,  $i \in I^*$ , and

$$\nabla f(x^*) + \sum_{i \in I} y_i^* \nabla g_i(x^*) = 0. \quad (7)$$

Then  $x^*$  is a local minimum point in problem (1).

**PROOF.** Take an arbitrary sequence of admissible points  $x^k \rightarrow x^*$ , and let  $s \in \mathbf{R}^n$ ,  $\|s\| = 1$ , be a limit point of the sequence of vectors  $(x^k - x^*)/\|x^k - x^*\|$ . Then from the differentiability of the  $g_i(x)$ ,  $i = r+1, \dots, m$ , at  $x^*$  it follows that

$$\begin{aligned} 0 &= g_i(x^k) = g_i(x^*) + (\nabla g_i(x^*), x^k - x^*) + o(x^k - x^*) \\ &= (\nabla g_i(x^*), x^k - x^*) + o(x^k - x^*). \end{aligned}$$

Dividing now by  $\|x^k - x^*\|$  and passing to the limit yields

$$(\nabla g_i(x^*), s) = 0, \quad i = r+1, \dots, m. \quad (8)$$

In exactly the same way we obtain

$$(\nabla g_i(x^*), s) \leq 0, \quad i \in I^*. \quad (9)$$

Now, we form the scalar product of (7) with  $s$  and use (8):

$$0 = (\nabla f(x^*), s) + \sum_{i \in I^*} y_i^* (\nabla g_i(x^*), s). \quad (10)$$

The quantity  $(\nabla g_i(x^*), s)$  cannot be zero for all  $i \in I^*$  since otherwise  $s \neq 0$  would be orthogonal to  $n$  linearly independent vectors in  $\mathbf{R}^n$ . Hence  $\phi(s) = \min_{i \in I^*} (\nabla g_i(x^*), s) < 0$  by (9). Introduce the set

$$\begin{aligned} S = \{s \in \mathbf{R}^n : \|s\| = 1, (\nabla g_i(x^*), s) = 0, i = r+1, \dots, m; \\ (\nabla g_i(x^*), s) \leq 0, i \in I^*\}. \end{aligned}$$

Then  $\max_{s \in S} \phi(s) = -\varepsilon < 0$  by the continuity and negativity of  $\phi(s)$  on  $S$  and the compactness of  $S$ .

Noting that  $y_i^* > 0, i \in I^*$ , from (10) we obtain

$$(\nabla f(x^*), s) = - \sum_{i \in I^*} y_i^* (\nabla g_i(x^*), s) \geq \varepsilon \min_{i \in I^*} y_i^* = 2\alpha > 0$$

for any  $s$  being limit points of  $(x^k - x^*)/\|x^k - x^*\|$  for admissible  $x^k$ ,  $x^k \rightarrow x^*$ . Since

$$f(x^k) \underset{k \rightarrow \infty}{\longrightarrow} f(x^*) + \|x^k - x^*\| (\nabla f(x^*), (x^k - x^*)/\|x^k - x^*\|) + o(x^k - x^*),$$

we find that

$$f(x^k) \geq f(x^*) + \alpha \|x^k - x^*\| \quad (11)$$

for all admissible  $x^k$  close enough to  $x^*$ . Thus  $x^*$  is a local minimum point.  $\square$

We have, in fact, proven (11), i.e.,  $x^*$  is a sharp minimum point (cf. Section 7.1).

For the elementary one-dimensional problem

$$\min f(x), \quad x \in \mathbf{R}^1, \quad x \geq 0 \quad (12)$$

Theorem 4 asserts that if  $f'(0) > 0$ , then  $x^* = 0$  is a minimum point. This fact is obvious geometrically (see Fig. 30).

It is sometimes impossible to make use of the sufficient first-order conditions, since Theorem 4 is applicable only when the number of active constraints is equal to the dimension of the space. For example, Theorem 4 is *a fortiori* invalid if the number of constraints is less than  $n$ .

**THEOREM 5** (sufficient second-order conditions). Let  $x^*$  be an admissible point and let the functions  $f(x)$ ,  $g_i(x)$ ,  $i = 1, \dots, m$ , be twice differentiable at  $x^*$ . Suppose that for some  $y^* \in \mathbf{R}^m$  condition (5) is satisfied and for every  $s \neq 0$  such that

$$\begin{aligned} (\nabla g_i(x^*), s) &= 0, \quad i \in I^*, \quad y_i^* > 0, \\ (\nabla g_i(x^*), s) &= 0, \quad i = r+1, \dots, m, \\ (\nabla g_i(x^*), s) &\geq 0, \quad i \in I^*, \quad y_i^* = 0, \end{aligned} \tag{13}$$

we have the inequality

$$(L''_{xx}(x^*, y^*)s, s) > 0. \tag{14}$$

Then  $x^*$  is a local minimum point in problem (1).  $\square$

### 9.2.3 Uniqueness and Stability of a Solution

**THEOREM 6.** Let the conditions of Theorem 4 or Theorem 5 hold. Then  $x^*$  is a locally unique solution.

Indeed, under the conditions of Theorem 4 we proved inequality (11), which indeed implies that the solution is unique. Similarly, one can show that under the conditions of Theorem 5, one has either the condition  $(\nabla f(x^*), s) > 0$  or the conditions  $(\nabla f(x^*), s) = 0$  and  $(\nabla^2 f(x^*)s, s) > 0$  hold, where  $s$  has the same significance as in Theorem 4. This easily implies the required assertion.  $\square$

The question of uniqueness of the Lagrange multipliers is also of interest to us.

**THEOREM 7.** Let the conditions of Theorem 2 be satisfied. Then the Lagrange multipliers are uniquely defined.  $\square$

This result is obvious because the converse contradicts regularity condition A. As was noted in Section 9.1 in connection with Slater's condition, regularity condition B does not imply uniqueness of the Lagrange multipliers.

Let us examine the stability of a solution toward perturbation of various kinds. The first result demonstrates the invariance of a solution under perturbations of the objective function if the sufficient first-order condition is satisfied.

**THEOREM 8.** Let the conditions of Theorem 4 be satisfied. Then a minimum point in the problem

$$\begin{aligned} & \underset{x \in Q}{\text{min}} [f(x) + \varepsilon f_1(x)] , \\ Q &= \{x: g_i(x) \leq 0, i = 1, \dots, r; g_i(x) = 0, i = r+1, \dots, m\} \end{aligned} \quad (15)$$

(where  $f_1(x)$  is differentiable at  $x^*$ ) for sufficiently small  $\varepsilon > 0$  coincides with  $x^*$ .

**PROOF.** By inequality (11)  $x^*$  is a sharp minimum point in problem (1). As in Theorem 8 of Section 7.1, it does not vary under small perturbations of the objective function.  $\square$

Next, we investigate a more general stability problem. Here and in what follows we shall need the following notion. We call the point  $x^* \in \mathbf{R}^n$  a nonsingular solution of problem (1) if

- (a) the functions  $f(x)$ ,  $g_i(x)$ ,  $i = 1, \dots, m$ , are twice differentiable in a neighborhood of  $x^*$ ;
- (b) the point  $x^*$  is admissible, the sets  $I^*$  and  $I$  are defined in (2);
- (c) regularity condition A is satisfied, i.e.,  $\nabla g_i(x^*)$ ,  $i \in I$ , are linearly independent;
- (d) the sufficient second-order condition and the strict complementarity condition are satisfied, i.e., we can find a vector  $y^*$  such that

$$L'_x(x^*, y^*) = 0, \quad y_i^* > 0, \quad i \in I^*; \quad y_i^* = 0, \quad i \in I \setminus I^*, \quad 1 \leq i \leq r, \quad (16) \quad \angle i$$

while for every  $s \neq 0$  such that

$$(\nabla g_i(x^*), s) = 0, \quad i \in I, \quad (17)$$

we have the inequality

$$(L''_{xx}(x^*, y^*)s, s) > 0. \quad (18)$$

The conditions of Theorem 5 are satisfied for such a point, hence it is indeed a minimum point. Furthermore, it is locally unique (see Theorem 6), and the Lagrange multipliers  $y^*$  are uniquely defined (see Theorem 7).

For a nonsingular minimum point the question concerning stability can be resolved in the most general formulation. Consider the perturbed problem

$$\begin{aligned} & \min (f(x) + \varepsilon_0 f_1(x)) , \\ & g_i(x) \leq \varepsilon_i, \quad i = 1, \dots, r , \\ & g_i(x) = \varepsilon_i, \quad i = r+1, \dots, m , \end{aligned} \quad (19)$$

where  $f_1(x)$  is a twice differentiable function in a neighborhood of  $x^*$ . Let  $\varepsilon$  denote the vector in  $\mathbf{R}^{m+1}$  with components  $\varepsilon_0, \dots, \varepsilon_m$ .

**THEOREM 9.** Let  $x^*$  be a nonsingular solution of (1). Then for sufficiently small  $\varepsilon$ , there exist a solution  $x_\varepsilon$  of problem (19) and the corresponding Lagrange multipliers  $y_\varepsilon$ . Here

$$x_\varepsilon \rightarrow x^*, \quad y_\varepsilon \rightarrow y^* \quad \text{as } \varepsilon \rightarrow 0, \quad (20)$$

where  $x_\varepsilon, y_\varepsilon$  are differentiable at 0 and for  $\phi(\varepsilon) = f(x_\varepsilon)$ ,

$$\frac{\partial \phi(0)}{\partial \varepsilon_0} = 0, \quad \frac{\partial \phi(0)}{\partial \varepsilon_i} = \sqrt{y_i^*}, \quad 1 \leq i \leq m. \quad (21)$$

Let us outline the proof. By the necessary extremum conditions at the point  $x_\varepsilon$  (if it exists and the set of active constraints at  $x_\varepsilon$  is the same as at  $x^*$ ) the following inequalities must hold:

$$\begin{aligned} \nabla f(x_\varepsilon) + \sum_{i \in I} y_\varepsilon i \nabla g_i(x_\varepsilon) + \varepsilon_0 \nabla f_1(x_\varepsilon) &= 0, \\ g_i(x_\varepsilon) &= \varepsilon_i, \quad i \in I. \end{aligned} \quad (22)$$

This system of equations in  $z = \{x, y_i, i \in I\}$  can be written in the form

$$R(z) = T(z)\varepsilon. \quad (23)$$

Here  $z^* = \{x^*, y_i^*, i \in I\}$  satisfies the equation  $R(z^*) = 0$ . The Jacobi matrix  $R'(z^*)$ , as is easily checked by direct calculation, has the form

$$R'(z^*) = \begin{pmatrix} L''_{xx}(x^*, y^*) & \nabla^T g(x^*) \\ \nabla g(x^*) & 0 \end{pmatrix}, \quad (24)$$

where  $\nabla g(x^*)$  is the matrix with rows  $\nabla g_i(x^*), i \in I$ . By Lemma 2 of Section 8.1 and condition (d) in the definition of a nonsingular minimum point, the matrix  $R'(z^*)$  is nonsingular. By Theorem 3 of Section 2.3, the system (23) has a solution  $z_\varepsilon$  and

$$z_\varepsilon = z^* - R'(z^*)^{-1} T(z^*) \varepsilon \neq o(\varepsilon). \quad (25)$$

It then follows that for sufficiently small  $\varepsilon$  we can find a solution  $x_\varepsilon, y_\varepsilon, i \in I$ , the system (23) that is differentiable with respect to  $\varepsilon$ . Here one has  $y_\varepsilon > 0, i \in I^*$ , since  $y_i^* > 0$ . It is possible to show that the sufficient second-order conditions are satisfied as well (since they are satisfied at  $x^*$  and also by continuity). The point  $x_\varepsilon$  is admissible since it satisfies

all the constraints (the constraints  $g_i(x) = \varepsilon_i$ ,  $i \in I$ , by definition of  $x_\varepsilon$ ; the constraints  $g_i(x) \leq \varepsilon_i$ ,  $i \notin I$ , since  $g_i(x^*) < 0$  for such  $i$  and by continuity). Thus  $x_\varepsilon$  is a solution of problem (19), with  $y_\varepsilon$  being the corresponding Lagrange multipliers.

Next, by the differentiability of  $x_\varepsilon$  we have

$$\begin{aligned}\phi(\varepsilon) - \phi(0) &= f(x_\varepsilon) - f(x^*) = (\nabla f(x^*), x_\varepsilon - x^*) + o(\varepsilon) \\ &= -\sum_{i=1}^m (y_i^* \nabla g_i(x^*), x_\varepsilon - x^*) + o(\varepsilon).\end{aligned}$$

Using formula (25) for  $z_\varepsilon - z^*$  and making the necessary calculations, we obtain (21) for  $\nabla \phi(0)$ .  $\square$

Expressions (21) make it possible to regard Lagrange multipliers to be constraint effect coefficients. Indeed, they show that the rate of change of the objective function under perturbation of any constraint is equal to the corresponding Lagrange multiplier. The greater the Lagrange multiplier, the greater the sensitivity with respect to the given equation. Conversely, for the inactive constraints one has  $y_i^* = 0$ , which implies that the solutions are insensitive to perturbation of these constraints.

### Exercises

3. Prove Theorem 8 differently, viz. show that the equation  $\nabla_x L_\varepsilon(x^*, y) = 0$  (where  $L_\varepsilon(x, y)$  is the Lagrange function for (15)) has a solution  $y_\varepsilon^*$  for sufficiently small  $\varepsilon$ , where  $y_\varepsilon^* > 0$ ,  $i \in I^*$ . Thus the sufficient extremum condition is satisfied at the point  $x^*$  for problem (15).

4. Prove that if all the requirements are satisfied in the definition of a nonsingular minimum, with the exception of (18), while the number of active constraints is equal to  $n$ , then (18) is satisfied too since  $s = 0$  follows from (17). Here the minimum is sharp, i.e.,  $f(x) \geq f(x^*) + \alpha \|x - x^*\|$ ,  $\alpha > 0$ , for all admissible  $x$  close enough to  $x^*$ .

## 9.3 CONVEX PROGRAMMING METHODS

Let us first consider the convex programming problem of the form

$$\begin{aligned}\min f(x), \quad x \in \mathbf{R}^n, \\ g_i(x) \leq 0, \quad i = 1, \dots, m,\end{aligned}\tag{1}$$

where  $f(x)$ ,  $g_i(x)$  are convex differentiable functions on  $\mathbf{R}^n$ . We are concerned only with the methods that are specific for this class of problems. It is worth noting that we may also apply to problem (1) methods designed

for more general problems (Section 9.4), as well as those designed for “simple” constraints (Chapter 7).

With regard to a classification of methods, we can repeat what was said in Section 8.2. However, in the given case, it is possible to classify the methods according to the principle as to whether they yield a sequence of admissible points (i.e., satisfying the constraints  $g_i(x) \leq 0$ ,  $i = 1, \dots, m$ ), or not. The former methods are referred to as *methods of feasible directions*.

### 9.3.1 Methods of Feasible Directions

Let  $x^k$  be some admissible point. We define for it the “ $\varepsilon$ -active” constraints, i.e., constraints which, to within the parameter  $\varepsilon_k \geq 0$ , turn into equalities:

$$I_k = \{i: g_i(x^k) \geq -\varepsilon_k\}. \quad (2)$$

We linearize the objective function and the  $\varepsilon$ -active constraints at the point  $x^k$  and find a direction  $s^k$  which is admissible for the linearized problem and at the same time leads to the most rapid decrease of the function. In other words, we take for  $s^k$  the solution of the auxiliary problem

$$\begin{aligned} & \max z, \\ & (\nabla f(x^k), s) \leq -z, \quad (\nabla g_i(x^k), s) \leq -z, \quad i \in I_k, \quad s \in S_k, \end{aligned} \quad (3)$$

where  $z \in \mathbf{R}^1$  is an additional variable and  $S_k$  is some “simple” bounded set which is introduced so that the problem be known to have a solution. For example, for  $S_k$  one can take some ball

$$S_k = \{s: \|s\|^2 \leq \rho_k^2\} \quad (4)$$

or cube

$$S_k = \{s: |s_i| \leq \rho_k, i = 1, \dots, n\}. \quad (5)$$

These methods are called the *normalizations N1* and *N2*, respectively. Next the step

$$x^{k+1} = x^k + \gamma_k s^k \quad (6)$$

is made, where the step size  $\gamma_k > 0$  is such that at the point  $x^{k+1}$  all the constraints not be violated and the  $f(x^{k+1})$  take on the smallest value. At the new point  $x^{k+1}$  the procedure is repeated.

The parameter  $\varepsilon_k$  can be governed in different ways; it is only necessary that  $\varepsilon_k > 0$ ,  $\varepsilon_k \rightarrow 0$ . If we take  $\varepsilon_k \equiv 0$  (i.e., taking into account only the active constraints at  $x^k$ ), then the method does not “sense” constraints, which hold at  $x^k$  “almost” like equalities (i.e.,  $|g_i(x^k)|$  is small), while their presence sharply constrains the step size from the point  $x^k$ . Therefore the method may “jam” in a neighborhood of a point that is not a solution.

One of the first concrete algorithms of the feasible-direction type was *Rosen's gradient projection method*: the step is made along the projection of the gradient  $\nabla f(x^k)$  onto the manifold defined by linearization of the active constraints. One should distinguish this method from the gradient projection method described in Section 7.2, with motion along the gradient followed by projection onto the admissible subset (this is different from the motion along the gradient projection onto the face of the admissible set; see Fig. 38, where the step of either method is shown for the case of linear constraints).

We omit theoretical results concerning the proof of the method of feasible directions for several reasons: (1) the proof is cumbersome (although not too complicated conceptually); (2) the rate of convergence of this method may be small even for well-posed problems—due to the parameter  $\varepsilon_k$  and the need to stay inside the admissible set; (3) the choice of the initial admissible point involves certain difficulties; and (4) at the present time, there are methods which are free from these drawbacks and at the same time are just as simple, and we shall describe one of them in the next subsection.

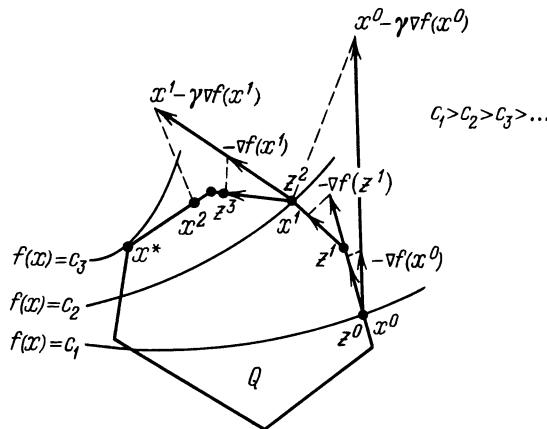


Fig. 38 The gradient-projection method (the points  $x^0, x^1, x^2, \dots$ ) and Rosen's method (the points  $z^0, z^1, z^2, \dots$ ).

### Exercises

1. Consider the gradient projection method and the conditional gradient method (Section 8.2) for problems with linear constraints as concrete examples of the method of feasible directions.

2. Show that problem (3), (4) is equivalent to the quadratic programming problem

$$\begin{aligned} \min & [-z + \lambda_k \|s\|^2], \\ (\nabla f(x^k), s) & \leq -z, \\ (\nabla g_i(x^k), s) & \leq -z, \\ i \in I_k, & \end{aligned}$$

for some  $\lambda_k \geq 0$ .

### 9.3.2 The Linearization Method

We linearize all the constraints at the point  $x^k$ , solve the auxiliary quadratic programming problem

$$\begin{aligned} \min & [(\nabla f(x^k), x - x^k) + (2\gamma)^{-1} \|x - x^k\|^2], \\ g_i(x^k) + (\nabla g_i(x^k), x - x^k) & \leq 0, \quad i = 1, \dots, m, \end{aligned} \tag{7}$$

and take its solution for  $x^{k+1}$  (cf. the analogous approach to the problem with equality constraints in Section 8.2).

**THEOREM 1.** Let the set  $X^*$  of solutions of problem (1) be nonempty, let the functions  $f(x)$ ,  $g_i(x)$  be differentiable, let their gradients satisfy a Lipschitz condition, and let Slater's condition hold. Then we can find a  $\bar{\gamma} > 0$  such that for  $0 < \gamma < \bar{\gamma}$  method (7) converges to a point  $x^* \in X^*$ . If, also,  $f(x)$  is strongly convex, then  $\|x - x^*\| \leq cq^k$ ,  $0 \leq q \leq 1$ .  $\square$

We omit the proof of this theorem since it is cumbersome, and since the local variant (under somewhat different assumptions) will be proved later for a more general problem (see Section 9.4). Note that if there are no constraints, method (7) becomes the gradient method of unconstrained minimization of  $f(x)$ , while Theorem 1 coincides with Theorem 1 of Section 6.1 and Theorem 2 of Section 1.4. Method (7) is rather close to the method of feasible directions with normalization N1, but does not generally coincide with it (cf. Exercise 2). In particular, the points derived in (7) do not need to be admissible.

In this variant of the method all the constraints are linearized. Actually, one may take into account only the “most violated” constraints. Exactly, given some  $\varepsilon > 0$ , we take

$$I_k = \{i: g_i(x^k) \geq \max_{1 \leq i \leq m} g_i(x^k)_+ - \varepsilon\} \tag{8}$$

and for  $x^{k+1}$  take the solution of the problem

$$\begin{aligned} \min & [(\nabla f(x^k), x - x^k) + (2\gamma)^{-1} \|x - x^k\|^2], \\ & g_i(x^k) + (\nabla g_i(x^k), x - x^k) \leq 0, \quad i \in I_k. \end{aligned} \tag{9}$$

For method (8), (9), assertions similar to Theorem 1 hold true. At the same time this modification of the method is more economical: at each step one solves a problem of smaller dimension than in the variant (7).

The way of choosing the constant parameter  $\gamma$  in (7) and (9) is not, of course, the only one possible. There exist constructive procedures for choosing  $\gamma_k$  similar to the algorithms for regulating the step size in the gradient method of unconstrained minimization (10) of Section 3.1.

Other computational schemes of method (7) or (9) can be used, too. Thus, transforming the objective function in (7), one can write method (7) in the form

$$\begin{aligned} x^{k+1} &= P_{Q_k}(x^k - \gamma \nabla f(x^k)), \\ Q_k &= \{x: g_i(x^k) + (\nabla g_i(x^k), x - x^k) \leq 0, i = 1, \dots, m\}. \end{aligned} \tag{10}$$

Hence, in particular, we see that for problem with linear constraints the method of linearization coincides with the gradient projection method (see Section 7.2). On the other hand, if we write out the dual problem for (9), then (see (12) in Section 10.4) it has the form

$$\min_{y_i \geq 0, i \in I_k} \left[ \frac{\gamma}{2} \left\| \nabla f(x^k) + \sum_{i \in I_k} y_i \nabla g_i(x^k) \right\|^2 - \sum_{i \in I_k} y_i g_i(x^k) \right]. \tag{11}$$

Let  $y_i^k$  denote its solution,  $i \in I_k$ , and let

$$x^{k+1} = x^k - \gamma \left( \nabla f(x^k) + \sum_{i \in I_k} y_i^k \nabla g_i(x^k) \right). \tag{12}$$

By virtue of (13) of Section 10.4, the point  $x^{k+1}$  is the same as in method (9).

The advantage of such an approach is the fact that auxiliary problem (11) reduces to finding the minimum of a quadratic function on  $\mathbb{R}_+^m$ , which can be found, say, by the conjugate gradient method in a finite number of steps (see Section 7.3).

### 9.3.3 Dual Methods

The scheme (11), (12) of the linearization method includes primal variable as well as dual variables. Many such methods are currently known. Apparently, the theoretically simplest scheme for updating the primal and

dual variables is the *Arrow-Hurwicz-Uzawa gradient method* (cf. (6) of Section 8.2):

$$\begin{aligned} x^{k+1} &= x^k - \gamma L'_x(x^k, y^k) = x^k - \gamma \left[ \nabla f(x^k) + \sum_{i=1}^m y_i^k \nabla g_i(x^k) \right], \\ y^{k+1} &= [y^k + \gamma L'_y(x^k, y^k)]_+ = [y^k + \gamma g(x^k)]_+. \end{aligned} \quad (13)$$

Here, for finding the saddle point of the function  $L(x, y)$  on  $\mathbf{R}^n \times \mathbf{R}_+^m$  (and problem (1) is equivalent to this; see Theorem 3 of Section 9.1) one makes a gradient step of minimization in  $x$  and a step of the gradient projection method in  $y$ . However, method (13) may not even converge. Consider an elementary example of a one-dimensional linear programming problem:

$$\begin{aligned} \min x, \\ x \in \mathbf{R}^1, \quad -x \leq 0, \end{aligned} \quad (14)$$

where  $x^* = 0$ ,  $y^* = 1$ , while method (13) has the form

$$x^{k+1} = x^k - \gamma(1 - y^k), \quad y^{k+1} = (y^k - \gamma x^k)_+. \quad (15)$$

The trajectory of such a process is shown in Figure 39. We do indeed find that for all  $\gamma$  the process does not converge (even to a neighborhood of  $x^*, y^*$ ).

If we assume that  $f(x)$  is strongly convex, then such effects are not observed. However, a rigorous justification of the method in this case is not an easy task. The standard technique of proving the method, based on a Lyapunov function of the form  $\|x - x^*\|^2 + \|y - y^*\|^2$  does not apply here (see Exercise 4). Only a weaker assertion can be proved in this manner, viz. that we can find a subsequence  $x^k$  converging to  $x^*$ . But such a result is not worth much, since it still not clear how to derive this subsequence. The corresponding subsequence  $y^k$  does not necessarily converge to  $y^*$ ; therefore  $L'_x(x^k, y^k)$  does not tend to 0 and cannot characterize the accuracy of the solution.

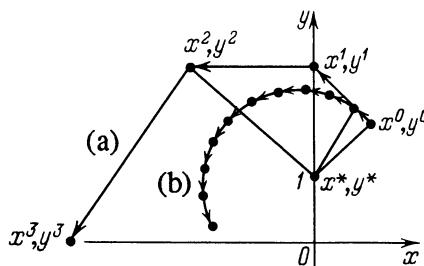


Fig. 39 Divergence of the gradient method for finding the saddle points: (a) large step; (b) small step.

Another variant of the Arrow-Hurwicz-Uzawa gradient method involves a complete minimization of  $L(x, y^k)$  in  $x$  instead of one gradient step (cf. (8) in Section 8.2):

$$x^{k+1} = \underset{x}{\operatorname{argmin}} L(x, y^k), \quad y^{k+1} = [y^k + \gamma g(x^{k+1})]_+. \quad (16)$$

This method may be interpreted as the gradient projection method for the dual problem. Indeed, as was shown in Section 9.1, the dual problem has the form

$$\max_{y \geq 0} \psi(y), \quad \psi(y) = \min_{x \in \mathbb{R}^n} L(x, y),$$

and hence if  $f(x)$  is strongly convex, then  $\psi(y)$  is concave, everywhere defined and differentiable, and the gradient of  $\psi(y)$  satisfies a Lipschitz condition and is equal to  $L'_y(y, y)$ , where  $x(y) = \underset{x}{\operatorname{argmin}} L(x, y)$  (Exercise 10 of Section 9.1). Hence one can write (16) as

$$y^{k+1} = [y^k + \gamma \nabla \psi(y^k)]_+, \quad (17)$$

and apply Theorem 1 of Section 7.2 on convergence of the gradient projection method.

The range of applicability of Lagrange multiplier methods can be enlarged and, also, the convergence rate can be increased if we pass from the usual Lagrangian over to the *augmented* Lagrangian. Let the function

$$M(x, y, K) = f(x) + \frac{1}{2K} \|y + K g(x)\|_+^2 - \frac{1}{2K} \|y\|^2, \quad (18)$$

where  $K$  is some parameter (cf. (14) of Section 8.1). We list its basic properties without proof.

**LEMMA 1.** Let  $f(x), g_i(x)$  be convex,  $K > 0$ . Then

(a)  $M(x, y, K)$  is convex in  $x$  and concave in  $y$ ;

(b)  $\lim_{K \rightarrow 0} M(x, y, K) = \begin{cases} L(x, y) & \text{if } y \geq 0, \\ -\infty & \text{otherwise;} \end{cases}$

(c) the sets  $X^* \times Y^*$  of saddle points of  $M(x, y, K)$  on  $\mathbb{R}^n \times \mathbb{R}^m$  and  $L(x, y)$  on  $\mathbb{R}^n \times \mathbb{R}_+^m$  coincide;

(d) if  $f(x), g_i(x)$  are differentiable, and  $X^* \times Y^*$  is nonempty, then

$$(M'_x(x, y, K), x - x^*) - (M'_y(x, y, K), y - y^*) \geq K \|M'_y(x, y, K)\|^2 \quad (19)$$

for all  $x \in \mathbb{R}^n, y \in \mathbb{R}^m, x^* \in X^*, y^* \in Y^*$ .  $\square$

Let us examine now the analog of method (13) in which  $L(x, y)$  is replaced by  $M(x, y, K)$  (cf. (9) of Section 8.2):

$$x^{k+1} = x^k - \gamma M'_x(x^k, y^k, K), \quad y^{k+1} = y^k + \gamma M'_y(x^k, y^k, K). \quad (20)$$

**THEOREM 2.** Let  $f(x)$ ,  $g_i(x)$  be convex (with  $f(x)$  being strongly convex), twice differentiable, and let their gradients and second derivatives satisfy a Lipschitz condition. Suppose that for the solution  $x^*$  and the Lagrange multipliers  $y^*$  the following conditions hold: the  $g_i(x^*)$  are linearly independent for  $i \in I^*$  and  $y_i^* > 0$  for  $i \in I^*$ , where  $I^* = \{i: g_i(x^*) = 0\}$ . Then for any  $x^0, y^0$  we can find a  $\bar{\gamma} > 0$  such that for  $0 < \gamma < \bar{\gamma}$  method (20) converges to  $x^*, y^*$  with the rate of geometric progression.

Here is a brief sketch of the proof. Introduce the Lyapunov function

$$V(x, y) = \|x - x^*\|^2 + \|y - y^*\|^2 - K(M(x, y, K) - M(x^*, y^*, K)). \quad (21)$$

Using inequality (19), one can show that

$$V(x^{k+1}, y^{k+1}) \leq V(x^k, y^k) - \lambda(\|M'_x(x^k, y^k, K)\|^2 + \|M'_y(x^k, y^k, K)\|^2),$$

$$\lambda > 0.$$

Therefore, the method converges. Hence for sufficiently large  $k$ , one has  $y_i^k \equiv 0$  for  $i \notin I^*$ . In this case the iteration process coincides with the method of the augmented Lagrangian for a problem with equality constraints

$$\begin{aligned} & \min f(x), \\ & g_i(x) = 0, \quad i \in I^*. \end{aligned} \quad (22)$$

For the latter equality, in Theorem 3 of Section 8.2 we proved the local convergence with the rate of geometric progression.  $\square$

An analog of (16) is the method in which the unconstrained minimum of the augmented Lagrangian is sought at each step (cf. (11) of Section 8.2):

$$x^{k+1} = \operatorname{argmin}_x M(x, y^k, K), \quad y^{k+1} = y^k + Kg(x^{k+1}). \quad (23)$$

**THEOREM 3.** Assume that the conditions of Theorem 2 are satisfied. Then for every  $y^0 \geq 0$  we can find a  $K$  such that for  $K > \bar{K}$  method (23) converges to the solution with the rate of geometric progression with ratio  $O(1/K)$ .

We restrict ourselves to sketching the proof. For a Lyapunov function it is appropriate to take  $V(y) = \|y - y^*\|^2$ . Then

$$V(y^{k+1}) \leq V(y^k) - \|y^{k+1} - y^k\|^2. \quad (24)$$

Therefore, the method converges. The local behavior of the method is the same as for the corresponding algorithm for solving problem (22). Using the result on the convergence rate of this algorithm (Theorem 4 in Section 8.2) yields the estimate of the convergence rate given in Theorem 3.  $\square$

One can regard method (23) to be one of the most effective approaches to solving convex programming problems. This method makes it possible to use powerful algorithms of unconstrained minimization for smooth functions (Chapter 3). At the same time, it has no drawbacks that are typical for the penalty-function method (see Section 9.3.4 below), viz. the penalty coefficient  $K$  does not grow, therefore the condition number of the auxiliary problems does not deteriorate from iteration to iteration. Finally, versus method (16) which is based on ordinary Lagrangians, method (23) has a significantly higher rate of convergence because the progression ratio can be made small by selection of  $K$ . One should, however, take into consideration that for a very large  $K$  the unconstrained minimization problems become too complicated (i.e., ill-posed).

To conclude, we mention one more method, in which the iterations over the primal and dual variables are the following:

$$\begin{aligned} y^k &= \underset{y_i \geq 0, i \in I_k}{\operatorname{argmin}} \left\| \nabla f(x^k) + \sum_{i \in I_k} y_i \nabla g_i(x^k) \right\|^2, \\ I_k &= \{i : g_i(x^k) \geq -\varepsilon\}, \\ x^{k+1} &= x^k - \gamma_k \left( \nabla f(x^k) + \sum_{i \in I_k} y_i^k \nabla g_i(x^k) \right). \end{aligned} \quad (25)$$

This is the method of *simultaneous solution of the primal and dual problems*. The meaning of this method is quite clear: for the point  $x^k$ , one finds the approximations  $y^k$  of the dual variables which minimize the residual in the extremum conditions. This approximation is used to refine the primal variables (i.e., one makes a step of the gradient method for minimizing  $L(x, y^k)$  in  $x$ ). Method (25) is close to (11), (12) of the linearization method. On the other hand, one can verify that (25) is obtained if one goes over to the dual problem in the method of feasible directions, with normalization N1, (3), (4), (6).

### Exercises

3. Show that in method (15), for (14) one has  $(x^{k+1} - x^*)^2 + (y^{k+1} - y^*)^2 > (x^k - x^*)^2 + (y^k - y^*)^2$  for any  $x^k, y^k > 0$  and any  $\gamma > 0$ .
4. Write method (13) for the problem in  $\mathbf{R}^1$ :  $\min x^2, x \leq 0$ , and verify that  $v_k = (x^k - x^*)^2 + (y^k - y^*)^2$  will not decrease monotonically for all  $x^k, y^k$  for an arbitrarily small  $\gamma > 0$ .

5. Prove Theorem 3, treating (23) as a gradient method for solving the following dual problem:  $\max_{y \in \mathbb{R}^m} \psi(y)$ , where  $\psi(y) = \inf_x M(x, y, K)$ .

### 9.3.4 Penalty Methods and Related Methods

The notions of the penalty-function method examined in Section 8.2 for problems with equality constraints are applicable as well to problems with inequality constraints. In the latter case this method is even more advantageous.

1. **P e n a l t y M e t h o d** (cf. (13) in Section 8.2):

$$\begin{aligned} x^k &= \operatorname{argmin}_x f_k(x), \\ f_k(x) &= f(x) + \frac{1}{2} K_k \sum_{i=1}^m g_i(x)_+^2, \quad K_k > 0, \quad K_k \rightarrow \infty. \end{aligned} \quad (26)$$

**THEOREM 4.** Let  $f(x)$ ,  $g_i(x)$  be convex and finite on  $\mathbb{R}^n$ , and let the set of solutions  $X^*$  of problem (1) be nonempty and bounded. Then  $f_k(x)$  is convex,  $X^* = \operatorname{Argmin}_{x \in \mathbb{R}^n} f_k(x) \neq \emptyset$ , sequence  $x^k$  is bounded and all its limit points belong to  $X^*$ , with  $f(x^k) \leq f^* = f(x^*)$ ,  $x^* \in X^*$ .  $\square$

The condition for  $X^*$  to be bounded is essential (see Exercise 6).

Under additional assumptions method (26) makes it possible to obtain the Lagrange multipliers, too. Indeed,

$$0 = \nabla f_k(x^k) = \nabla f(x^k) + K \sum_{i=1}^m g_i(x^k)_+ \nabla g_i(x^k).$$

Since  $x^k \rightarrow x^*$ , and at  $x^*$  we have

$$\nabla f(x^*) + \sum_{i=1}^m y_i^* \nabla g_i(x^*) = 0,$$

 then necessarily  $K g_i(x^k) \rightarrow y_i^*$ .

2. **B a r r i e r M e t h o d:**

$$\begin{aligned} x^k &= \operatorname{argmin}_x f_k(x), \\ f_k(x) &= f(x) - \varepsilon_k \sum_{i=1}^m \frac{1}{g_i(x)}, \quad \varepsilon_k > 0, \quad \varepsilon_k \rightarrow 0. \end{aligned} \quad (27)$$

Here the minimum is taken over  $x$  for which  $g_i(x)$ ,  $i = 1, \dots, m$ . This method is also called the *interior penalty method*, since, unlike (26), which is the

*exterior penalty method*, the penalty in (27) is not equal to 0 even for admissible points and it increases as it approaches the boundary from the inside.

**THEOREM 5.** Let  $f(x)$ ,  $g_i(x)$  be convex and finite, let the set of solutions  $X^*$  of problem (1) be nonempty and bounded, and let Slater's condition be satisfied. Then in method (27),  $X_k^* = \operatorname{Argmin}_x f_k(x) \neq \emptyset$ , the  $f_k(x)$  are convex, the sequence  $x^k$  is bounded and all its limit points belong to  $X^*$ , with  $f(x^k) \geq f^* = f(x^*)$ ,  $x^* \in X^*$ .  $\square$

In method (27) one can also obtain approximations for the Lagrange multipliers (Exercise 7). The function  $1/g_i(x)$  can be replaced by other functions, say, by  $\log(-g_i(x))$ .

**3. Penalty Shifting Method.** One can regulate the penalty not only by choosing the penalty coefficients but also by shifting the penalty level:

$$\begin{aligned} x^k &= \operatorname{Argmin}_x f_k(x), \\ f_k(x) &= f(x) + K \sum_{i=1}^m (g_i(x) + \lambda_i^k)_+^2, \end{aligned} \quad (28)$$

where it is necessary to raise the levels  $\lambda_i^k$  if  $x^k$  is inadmissible and to lower them if  $x^k$  is admissible. For example, one can take

$$\lambda_i^{k+1} = \lambda_i^k + g_i(x^k), \quad i = 1, \dots, m. \quad (29)$$

If we substitute for  $y_i^k = K\lambda_i^k$ , we see that method (28), (29) turns into method (23). Therefore, the method of the augmented Lagrangian (23) can be interpreted as the *penalty shifting method*.

**4. Method of Selecting  $f^*$ .** If  $f^* = \min \{f(x) : g_i(x) \leq 0, i = 1, \dots, m\}$  were known, then problem (1) would be equivalent to that of minimizing the function  $(f(x) - f^*)^2 + \|g(x)_+\|^2$ . One can select  $f^*$  recursively by solving the problem

$$x^k = \operatorname{Argmin}_x f_k(x), \quad f_k(x) = (f(x) - f_k)_+^2 + \sum_{i=1}^m g_i(x)_+^2. \quad (30)$$

Clearly, if  $f_k(x^k) > 0$ , then the value of  $f_k$  should be increased. One of the possible rules for updating  $f_k$  is

$$f_{k+1} = f_k + f_k(x^k)/(f(x^k) - f_k). \quad (31)$$

**THEOREM 6.** Let  $f(x)$ ,  $g_i(x)$  be convex and finite, let  $x^*$  be a regular minimum point of problem (1) and let  $f_0 < f^*$ . Then in method (30), (31),  $x^k \rightarrow x^*$ ,  $f_k \leq f^*$  and

$$g_i(x^k)_+ / (f(x^k) - f_k) \rightarrow y_i^*, \quad i = 1, \dots, m. \quad \square \quad (32)$$

Let us compare the methods described above. Figure 40 shows the forms of the functions  $f_k(x)$  for each method for the elementary problem  $\min x$ ,  $x \in \mathbb{R}^1$ ,  $-x \leq 0$ .

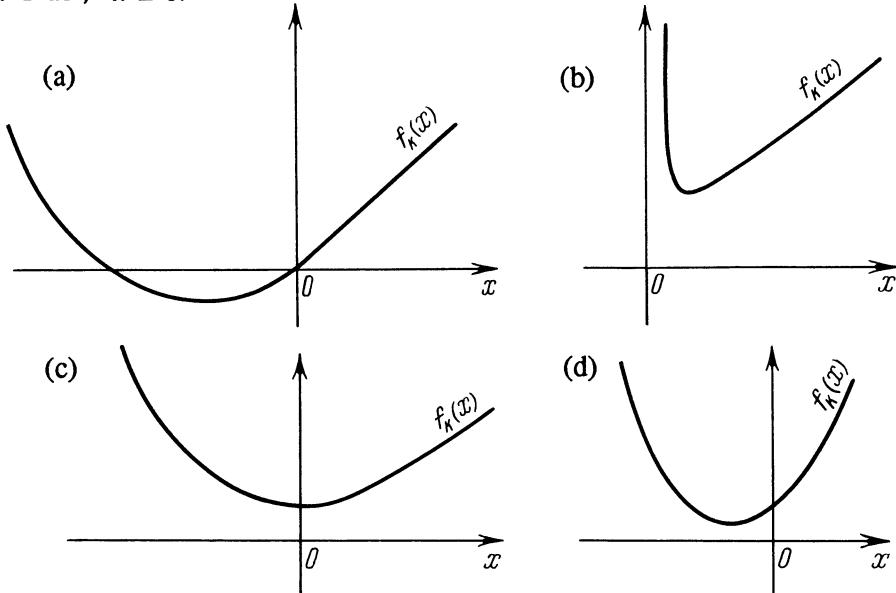


Fig. 40 The auxiliary problems of unconstrained minimization:  
 (a) the penalty method; (b) the barrier method; (c) the penalty shifting method; (d) the method for selecting  $f^*$ .

minor

The penalty function method (26) converges under ~~marginal~~ assumptions. It reduces to unconstrained minimization of a differentiable convex function on the entire  $\mathbb{R}^n$ , and it does not require that an initial admissible point be found. The fact that the points  $x^k$  do not satisfy the constraints is a disadvantage of this method. It can, however, be easily eliminated, viz. if a point  $x^0$  is known, such that  $g_i(x^0) < 0$ ,  $i = 1, \dots, m$ , then one can take

$$\bar{x}^k = \lambda_k x^0 + (1 - \lambda_k)x^k,$$

$$\lambda_k = \min \{\lambda: g_i(\lambda x^0 + (1 - \lambda)x^k) \leq 0, i = 1, \dots, m\}.$$

Then the  $\bar{x}^k$  are admissible and they converge to the solution. In this case  $f(\bar{x}^k) \geq f^* \geq f(x^k)$ , i.e., we obtain two-sided bounds of the solution ac-

curacy. A more serious drawback of all variants of the penalty method (cf. Section 8.2) is the fact that the function  $f_k(x)$  is ill-conditioned for large  $K$ , which substantially complicates solution of the auxiliary unconstrained minimization problem in (26).

The only advantage that the barrier method has over the penalty method is, perhaps, the fact that the function  $f_k(x)$  in (27) is smoother than in (26) (where it is once differentiable). At the same time, for (27) to converge, Slater's condition is required (otherwise the method becomes meaningless); one needs to know an initial point  $x^0$  such that  $g_i(x^0) < 0$ ,  $i = 1, \dots, m$ ; while minimizing the process, one should prevent this point from leaving the admissible region. The latter circumstance is especially unfavorable and does not permit the application of standard unconstrained minimization techniques. However, they can be modified by suitably changing the procedure for determining the step size. Finally, the same characteristics of the ill-conditionality is typical for method (27): the function  $f_k(x)$  behaves essentially different in different directions (it grows sharply as it approaches the boundary and hardly varies at all as it moves along the boundary).

As we have seen, the penalty shifting method (28), (29) coincides with the method of augmented Lagrangian and possesses all its merits: the penalty coefficient  $K$  is not required to grow and hence the method does not lead to ill-posed problems; it is possible to raise the convergence rate by appropriately choosing  $K$ , and so on. Its disadvantages include the more stringent conditions required for the method to converge.

To sum up, the method of choosing  $f^*$  (30), (31) has a number of attractive features: it contains no coefficients tending to 0 or  $\infty$ ,  $f_k(x)$  is everywhere defined, and so on. However, it has the disadvantage of ill-conditionality even for regular problems (see Exercise 8). Moreover, it requires that the auxiliary minimization problems be solved with high accuracy since it involves small quantities  $f(x^k) - f_k$ . For example, if  $f_k > f^*$ , then  $\min_x f_k(x) = 0$ ; however, due to unavoidable errors, the point  $x^k$  is found approximately, and it may turn out that  $f(x^k) > f_k$  and the value of  $f_k$  becomes greater (whereas one needs, in fact, to decrease it).

## Exercises

6. Consider the minimization problem in  $\mathbf{R}^2$ :  $\min x_1, -x_1 + 1 \leq 0, \rho^2(x, Q) \leq 0$ , where  $Q = \{x: x_2 \geq x_1^{-1}, x_1 > 0\}$ . Here  $X^* = \{x: x_1 = 1, x_2 \geq 1\}$ , the function  $g_2(x) = \rho^2(x, Q)$  is convex (being the square of the distance to a convex set  $Q$ , see Exercise 2 in Section 5.1). Show that for  $0 < K \leq 1$  the function  $f_k(x)$  in (26) has no minimum point for this problem.
7. Prove that in (27) for a nonsingular minimum point one has  $\varepsilon_k/g_i(x^k)^2 \rightarrow y_i^*$ .

- 8.** Consider the problem in  $\mathbf{R}^2$ :  $\min x_1, -x_1 \leq 0, (x_1+1)^2 + x_2^2 - 1 \leq 0; x^* = 0, f^* = 0$ . Show that the function  $\phi(x) = (f(x) - f^*)_+^2 + g_1(x)_+^2 + g_2(x)_+^2$  (in method (30), (31) for  $f_k = f^*$ ) has a singular minimum at the point  $x^*$  (in particular,  $\partial^2\phi(0)/\partial x_2^2 = 0$ ).

### 9.3.5 Methods for Nonsmooth Problems

We shall be examining the problem

$$\begin{aligned} & \min f(x), \\ & g_i(x) \leq 0, \quad i = 1, \dots, m, \quad x \in Q, \end{aligned} \tag{33}$$

where  $Q$  is a closed convex set in  $\mathbf{R}^n$ , the functions  $f(x)$ ,  $g_i(x)$  are convex and finite on  $\mathbf{R}^n$ . Here it is assumed that  $Q$  is “simple” in the sense of the arguments in Chapter 7, while  $f(x)$  and  $g_i(x)$  are generally nondifferentiable, and at an arbitrary point  $x \in Q$  one can compute their subgradients  $\partial f(x)$ ,  $\partial g_i(x)$ . Clearly, the  $m$  constraints  $g_i(x) \leq 0$  are equivalent to the one constraint

$$g(x) \leq 0, \quad g(x) = \max_{1 \leq i \leq m} g_i(x), \tag{34}$$

where  $g(x)$  is convex and by Lemma 11 of Section 5.1

$$\begin{aligned} \partial g(x) &= \left\{ \sum_{i \in I(x)} \lambda_i \partial g_i(x), \lambda \geq 0, \sum_{i \in I(x)} \lambda_i = 1 \right\}, \\ I(x) &= \{i: g_i(x) = g(x)\}. \end{aligned}$$

Consider the following extension of the *subgradient projection method*:

$$\begin{aligned} x^{k+1} &= P_Q(x^k - \gamma_k s^k), \\ \gamma_k &\rightarrow 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty, \\ s^k &= \begin{cases} \partial f(x^k) & \text{if } g(x^k) \leq 0, \\ \partial g(x^k) & \text{if } g(x^k) > 0. \end{cases} \end{aligned} \tag{35}$$

In other words, the step is made along the projection of the subgradient of the objective function if the point  $x^k$  is admissible, and along the projection of the subgradient of the violated constraints otherwise.

**THEOREM 7.** Let the set  $Q$  be bounded and let there exist an  $x^0 \in Q$ ,  $g_i(x^0) < 0$ ,  $i = 1, \dots, m$ . Then all the limit points of the sequence  $x^k$  in method (35) are solutions to problem (33).

**PROOF.** By our assumption, the  $f(x)$ ,  $g_i(x)$  are continuous on  $Q$ , the admissible set is closed and bounded, and therefore a solution exists. Let  $x^*$  be any solution. Take  $\varepsilon > 0$  and consider  $S_\varepsilon = \{x \in Q : f(x) \leq f(x^*) + \varepsilon, g(x) \leq 0\}$ . Then on the segment  $[x^0, x^*]$  there is a point  $\bar{x} \in S_\varepsilon$  such that  $f(\bar{x}) \leq f(x^*) + \delta$ ,  $g(\bar{x}) \leq -\delta$  for some  $0 < \delta < \varepsilon$ . Let us now estimate the distance from  $x^{k+1}$  to  $\bar{x}$ . We have

$$\begin{aligned}\|x^{k+1} - \bar{x}\|^2 &= \|P_Q(x^k - \gamma_k s^k) - \bar{x}\|^2 \leq \|x^k - \gamma_k s^k - \bar{x}\|^2 \\ &= \|x^k - \bar{x}\|^2 - 2\gamma_k(s^k, x^k - \bar{x}) + \gamma_k^2 \|s^k\|^2.\end{aligned}$$

Since the subgradient of the continuous function  $f(x)$  or  $g(x)$  is bounded on the bounded set  $Q$  (see Lemma 8 in Section 5.1), then  $\|s^k\|^2 \leq c$ . If  $x^k \notin S_\varepsilon$ , then either  $g(x^k) \leq 0$ ,  $f(x^k) \geq f(x^*) + \varepsilon$ , or  $g(x^k) > 0$ ,  $f(x^k) \leq f(x^*) + \varepsilon$ . In the former case,  $s^k = \partial f(x^k)$ ,  $(s^k, x^k - \bar{x}) = (\partial f(x^k), x^k - \bar{x}) \geq f(x^k) - f(\bar{x}) \geq f(x^*) + \varepsilon - f(x^*) - \delta = \varepsilon - \delta$ . In the latter case,  $s^k = \partial g(x^k)$ ,  $(s^k, x^k - \bar{x}) = (\partial g(x^k), x^k - \bar{x}) \geq g(x^k) - g(\bar{x}) \geq -g(\bar{x}) \geq \delta$ .

Thus, if  $x^k \notin S_\varepsilon$ , then

$$\|x^{k+1} - \bar{x}\|^2 \leq \|x^k - \bar{x}\|^2 - 2\alpha\gamma_k + \gamma_k^2 c, \quad \alpha = \min \{\delta, \varepsilon - \delta\} > 0.$$

This implies in the usual way (see, for instance, the proof of Theorem 1 in Section 5.3) that this inequality cannot be satisfied for all  $k$ . Thus, for any  $\varepsilon > 0$  there is an  $x^k \in S_\varepsilon$ . Since  $\bigcap_{\varepsilon > 0} S_\varepsilon = X^*$ , where  $X^*$  is the set of solutions of (42), all the limit points of  $x^k$  belong to  $X^*$ .  $\square$

Another way of applying the subgradient method to problem (33) is to reduce it to an unconstrained minimization problem by using the following lemma.

**LEMMA 2.** Let Slater's condition be satisfied and let the set of solutions  $X^*$  of problem (33) be nonempty. Then  $X^*$  coincides with the set of minimum points of the function

$$\phi(x) = f(x) + K \sum_{i=1}^m g_i(x)_+ \tag{36}$$

on  $Q$  for all  $K > \bar{K} = \max_{1 \leq i \leq m} y_i^*$ , where  $y_i^*$  are the Lagrange multipliers.

This result follows directly from the Kuhn-Tucker theorem and the formula for  $\partial\phi(x)$  (based on Lemmas 10 and 11 in Section 5.1).  $\square$

Thus one can choose a sufficiently large  $K$  and apply the subgradient projection method for minimizing  $\phi(x)$ :

$$\begin{aligned} x^{k+1} &= P_Q(x^k - \gamma_k \partial\phi(x^k)) = P_Q \left[ x^k - \gamma_k \left( \partial f(x^k) + K \sum_{i=1}^m \mu_{ik} \partial g_i(x^k) \right) \right], \\ \mu_{ik} &= \begin{cases} 0, & g_i(x^k) \leq 0, \\ 1, & g_i(x^k) > 0. \end{cases} \end{aligned} \quad (37)$$

Since the lower bound  $\bar{K}$  for  $K$  is usually unknown, one can make this parameter variable by increasing it at each step:

$$x^{k+1} = P_Q \left[ x^k - \gamma_k \left( \partial f(x^k) + K_k \sum_{i=1}^m \mu_{ik} \partial g_i(x^k) \right) \right], \quad K_k \rightarrow \infty. \quad (38)$$

The above described methods converge, naturally, no faster than the subgradient methods of unconstrained minimization. In this connection, for large  $K$  the function  $\phi(x)$  in (36) has strongly elongated level lines, and is hard to minimize. Hence one cannot count on such methods to be efficient.

The techniques for increasing the rate of convergence in the subgradient methods, just as in an unconstrained minimization problem, suggest that in the  $k$ th iteration the information obtained in the previous iterations should be used.

The most straightforward approach consists in constructing a piecewise-linear approximation of the objective function and of the constraints. Suppose the subgradients have already been computed at the points  $x^1, \dots, x^k$ . Then  $f(x)$  can be replaced by the expression  $\max_{1 \leq j \leq k} [f(x^j) + (\partial f(x^j), x - x^j)]$ , and one may proceed similarly with the  $g_i(x)$ . Then the initial problem is approximated by the following problem:

$$\begin{aligned} \min z, \\ f(x^j) + (\partial f(x^j), x - x^j) \leq z, \quad j = 1, \dots, k, \\ g_i(x^j) + (\partial g_i(x^j), x - x^j) \leq 0, \quad i = 1, \dots, m, j = 1, \dots, k, \quad x \in Q, \end{aligned} \quad (39)$$

which is a linear programming problem (if  $Q$  is a polyhedron). Its solution can be taken for  $x^{k+1}$ . It is not hard to see that provided there are no constraints  $g_i(x) \leq 0$  this method coincides with the cutting-plane method described in Section 5.4. One can reduce the number of constraints in the auxiliary problem (39) if all the constraints are replaced by the single constraint  $g(x) \leq 0$ . It is convenient to take  $g(x) = \max_{1 \leq i \leq m} g_i(x)$  since a calculation of the subgradient of  $g(x)$  requires a calculation of only one subgradient of  $g_i(x)$  (i.e.,  $\partial g(x) = \partial g_i(x)$ ,  $j = \arg \max_{1 \leq i \leq m} g_i(x)$ ).

There are also other ways of constructing piecewise linear approximations of a convex programming problem. In addition, some methods derived in Section 5.4 for unconstrained minimization problems, i.e., the method of Chebyshev centers, the center-of-gravity method, the space extension method, among others, carry over to constrained problems.

Moreover, the methods described above which involve dual variables extend easily to the nonsmooth case: one needs only to replace the gradient method by the subgradient projection method. In particular, the elementary algorithm (13) takes on the form

$$\begin{aligned} x^{k+1} &= P_Q [x^k - \gamma_k \partial_x L(x^k, y^k)] = P_Q \left[ x^k - \gamma_k \left( \partial f(x^k) + \sum_{i=1}^m y_i^k \partial g_i(x^k) \right) \right], \\ y^{k+1} &= [y^k + \gamma_k \partial_y L(x^k, y^k)]_+ = [y^k + \gamma_k g(x^k)]_+, \\ &\quad \gamma_k \rightarrow 0, \sum_{k=0}^{\infty} \gamma_k = \infty. \end{aligned} \quad (40)$$

Using the same example (14), it is not hard to check that method (49), without additional assumptions such as strict convexity of  $f(x)$ , does not converge. The question as to the convergence of the method has not yet been resolved.

If the minimum of  $L(x, y)$  for  $x \in Q$  for every  $y \in \mathbf{R}_+^m$  exists and is easily found, then it is possible to apply an analog of method (16):

$$\begin{aligned} x^{k+1} &= \underset{x \in Q}{\operatorname{argmin}} L(x, y^k), \\ y^{k+1} &= [y^k + \gamma_k g(y^k)]_+, \quad \gamma_k \rightarrow 0, \sum_{k=0}^{\infty} \gamma_k = \infty. \end{aligned} \quad (41)$$

Interpreting it as a subgradient method for solving the dual problem

$$\max_{y \geq 0} \psi(y), \quad \psi(y) = \underset{x \in Q}{\operatorname{min}} L(x, y) \quad (42)$$

makes it possible to obtain results on convergence.

Unfortunately, in nonsmooth problems one is rarely able to find the minimum of  $L(x, y)$  in  $x$  in explicit form, which makes method (41) difficult to use (as well as its analog, in which the  $L(x, y)$  is replaced by  $M(x, y, K)$ ).

## Exercises

9. Investigate the modification of method (35) in which for  $g(x)$  in (34) one takes  $g(x) = \sum_{i=1}^m g_i(x)_+$ .

- 10.** Consider other rules for choosing the step size in method (35) similar to those described in Section 5.3. Examine their convergence and check in which particular cases the condition for  $Q$  to be bounded may be dropped.  
**11.** Analyze the method of type (35) in which  $s^k = \partial f(x^k)$  for  $g(x^k) \leq -\varepsilon_k$ ,  $s^k = \partial g(x^k)$  for  $g(x^k) > -\varepsilon_k$ ,  $\varepsilon_k \geq 0$ ,  $\varepsilon_k \rightarrow 0$ .

### 9.3.6 Summary

In Section 8.3 we pointed out the difficulties that may arise in solving problems with equality constraints. Let us see now to what extent these difficulties are characteristic of convex programming methods.

To begin, the multimodality issue that poses a major “hindrance” in nonconvex problems does not come up in convex problems. Indeed, all the auxiliary minimization problems considered above (for the linearization method, the dual methods, the penalty method, among others) are convex, hence the local minima coincide with global minima.

Furthermore, assertions concerning the convergence of convex programming methods are global. Therefore there is no need to have good initial approximations, neither one has to expect a divergence of the methods. In this respect, the noise poses no special threat: if in the problems described in Chapter 8 noise could lead the method out of the convergence domain, this situation does not occur in the convex case. We shall not analyze in detail the influence of noise of various kinds—such analysis is similar in many ways to the one in Chapter 4 and in Sections 5.5 and 7.4. We restrict ourselves to a brief remark.

It may occur in nonconvex problems that the auxiliary minimization problems in some method have no solution, e.g., if the constraints are contradictory, even if in the initial problem the minimum has been attained. In convex programming this is not possible. For instance, the quadratic programming problem (7) in the linearization method always has a solution since any admissible point (1) satisfies the constraints in (7). If there are contradictory constraints in the auxiliary problem, the same situation is observed in the initial problem.

To conclude, convex programming problems are free from many complications which are typical for more general problems.

## 9.4 NONLINEAR PROGRAMMING METHODS

Let consider now the same problem as that in Section 9.2:

$$\begin{aligned} \min f(x), \quad & x \in \mathbf{R}^n, \\ g_i(x) \leq 0, \quad & i = 1, \dots, r, \\ g_i(x) = 0, \quad & i = r+1, \dots, m, \end{aligned} \tag{1}$$

where  $f(x)$ ,  $g_i(x)$  are differentiable functions. Methods for solving this problem and the results on convergence are close to those in Section 8.2 and Section 9.3, where we investigated the special cases of problem (1). Hence we shall be brief.

#### 9.4.1 The Linearization Method

The idea of this method, i.e., linearization of the objective function and of the constraints, is well known. For the general problem (1) the method takes on the form

$$\begin{aligned} \min & [(\nabla f(x^k), x - x^k) + (2\gamma)^{-1} \|x - x^k\|^2], \\ g_i(x^k) + (\nabla g_i(x^k), x - x^k) & \leq 0, \quad i \in I_k^*, \\ g_i(x^k) + (\nabla g_i(x^k), x - x^k) & = 0, \quad i = r+1, \dots, m, \\ I_k^* & = \{i: 1 \leq i \leq r, g_i(x^k) \geq -\varepsilon\}, \quad \varepsilon > 0. \end{aligned} \tag{2}$$

For  $x^{k+1}$  one takes the solution of this problem, and the process is repeated. Note that not all of the constraints are linearized, just the “ $\varepsilon$ -active” ones, i.e., for  $i \in I_k^*$ . Instead of the quadratic programming problem (2) one can introduce its dual problem (see (12) of Section 10.4):

$$\begin{aligned} \min_{y \in S_k} & \left[ \gamma \left\| \nabla f(x^k) + \sum_{i=1}^m y_i \nabla g_i(x^k) \right\|^2 - \sum_{i=1}^m y_i g_i(x^k) \right]. \\ S_k & = \{y \in \mathbb{R}^m: y_i \geq 0, i \in I_k^*; y_i = 0, i \in I_k^*, 1 \leq i \leq r\}. \end{aligned} \tag{3}$$

This problem of minimizing a quadratic function, provided a subset of the variables is nonnegative, can be solved by the conjugate-gradient method (Section 7.3) in a finite number of steps. One can find its solution  $y_i^k$ , and compute the new approximation for the primal variables (see (13) in Section 10.4):

$$x^{k+1} = x^k - \gamma \left( \nabla f(x^k) + \sum_{i=1}^m y_i^k \nabla g_i(x^k) \right). \tag{4}$$

Since we make no assumptions concerning the convexity, it is appropriate that the results on the convergence of the method are of the local nature. Moreover, it is required that the solution be nonsingular (see Section 9.2).

**THEOREM 1.** Let  $x^*$  be a nonsingular minimum point of (1). Then we can find  $\bar{\gamma} > 0$ ,  $\bar{\varepsilon} > 0$ , such that for  $0 < \gamma < \bar{\gamma}$ ,  $0 < \varepsilon < \bar{\varepsilon}$ , method (2) converges locally to a solution with the rate of geometric progression.

We sketch the proof. It is clear that for sufficiently small  $\varepsilon > 0$  and  $x^k$  sufficiently close to  $x^*$ , one has  $I_k^* = I^*$ . Furthermore, under the same assumptions it is possible to show that in problem (2) one has  $g_i(x^k) + (\nabla g_i(x^k), x^{k+1} - x^k) = 0$ ,  $i \in I_k^* = I^*$ , i.e., the constraints which are active for  $x^*$  in (1) are also active for a solution of (2). Thus the local method coincides with the linearization method for problems with equality constraints which are active at the point  $x^*$ :

$$\begin{aligned} \min f(x) , \\ g_i(x) = 0 , \quad i \in I . \end{aligned} \tag{5}$$

But convergence of the linearization method with equality constraints was proved earlier (see Theorem 1 of Section 8.2).  $\square$

It turns out that for sharp minimum problems the linearization method converges locally with the quadratic rate.

**THEOREM 2.** Let  $x^*$  be a nonsingular minimum point and let the number of active constraints be equal to  $n$ . Then for sufficiently small  $\gamma > 0$ ,  $\varepsilon > 0$ , method (2) converges locally to a solution with the quadratic rate.  $\square$

#### 9.4.2 Newton-like and Quasi-Newton Methods

If the second derivatives of all the functions  $f(x)$ ,  $g_i(x)$  are available, and if it is not too difficult to compute them, one can apply the following *analog of Newton's method*. At the point  $x^k$  the constraints are linearized, the Lagrange function is approximated by a quadratic function and the auxiliary quadratic programming problem

$$\begin{aligned} \min & [(L'_x(x^k, y^k), x - x^k) + (L''_{xx}(x^k, y^k)(x - x^k), x - x^k)/2] , \\ & g_i(x^k) + (\nabla g_i(x^k), x - x^k) \leq 0 , \quad i \in I_k^* , \\ & g_i(x^k) + (\nabla g_i(x^k), x - x^k) = 0 , \quad i = r+1, \dots, m , \\ & I_k^* = \{i: 1 \leq i \leq r, g_i(x^k) \geq -\varepsilon\} , \quad \varepsilon > 0 , \end{aligned} \tag{6}$$

is solved. For its solution one takes  $x^{k+1}$ , for the Lagrange multipliers one takes  $y^{k+1}$  (for  $i \notin I_k^*$  one takes  $y_i^{k+1} = 0$ ). For the special case with the equality constraints only, this method was considered earlier (see (19) in Section 8.2).

**THEOREM 3.** Let  $x^*$  be a nonsingular minimum point and let the second derivatives of the functions  $f(x)$ ,  $g_i(x)$  satisfy a Lipschitz condition in a neighborhood of  $x^*$ . Then for sufficiently small  $\varepsilon > 0$  method (6) converges locally to  $x^*$ ,  $y^*$  with quadratic rate.

This theorem can be proven if we show first that method (6) coincides locally with Newton's method for solving problem (5) and use the fact that the latter converges (Theorem 9 in Section 8.2).  $\square$

Method (6) is generally very laborious and requires that the second derivatives of the objective function as well as the constraints be computed; it also requires that the quadratic programming problem be solved at each step. One can eliminate the first drawback by passing to quasi-Newton methods in the same way as for unconstrained minimization (Section 3.3). Namely, instead of (6) one needs to solve at each step the auxiliary problem

$$\begin{aligned} \min & [(L'_x(x^k, y^k), x - x^k) + (H_k(x - x^k), x - x^k)/2], \\ g_i(x^k) + (\nabla g_i(x^k), x - x^k) & \leq 0, \quad i \in I_k^*, \\ g_i(x^k) + (\nabla g_i(x^k), x - x^k) & = 0, \quad i = r+1, \dots, m, \\ I_k^* & = \{i: 1 \leq i \leq r, g_i(x^k) \geq -\varepsilon\}, \quad \varepsilon > 0, \end{aligned} \tag{7}$$

where the matrix  $H_k$  is the approximation for  $L''_{xx}(x^k, y^k)$  and has the form

$$H_k = F_k + \sum_{i \in I_k} y_i^k G_i^k, \quad I_k = \{i: i \in I_k^* \cup i = r+1, \dots, m\}, \tag{8}$$

$F_k$ ,  $G_i^k$  being the approximations for  $\nabla^2 f(x^k)$ ,  $\nabla^2 g_i(x^k)$  constructed from the values of the gradients of these functions at the preceding points by means of recurrence relations (see Section 3.3).

A close variant of the method involves solving in the  $k$ th iteration the auxiliary minimization problem under linear constraints (cf. (24) in Section 8.2):

$$\begin{aligned} \min & L(x, y^k), \\ g_i(x^k) + (\nabla g_i(x^k), x - x^k) & \leq 0, \quad i \in I_k, \\ g_i(x^k) + (\nabla g_i(x^k), x - x^k) & = 0, \quad i = r+1, \dots, m. \end{aligned} \tag{9}$$

The rule for updating the  $y^k$  for method (9) is the same as before. For (9) the same convergence results are valid as for Newton's method (6).

The possible nonconvexity of the auxiliary problem is a major disadvantage of method (6) and similar methods. The matrix  $L''_{xx}(x^k, y^k)$  is generally not positive definite even for  $x^k, y^k$  arbitrarily close to  $x^*$ ,  $y^*$ . Hence the objective function can be nonconvex, and the whole auxiliary problem (6) can be multimodal, which substantially complicates the solution of the problem. In this case it is more appropriate to use methods such as (7) with positive definite matrices  $H_k$ . For example, one can begin with  $H_0 = I$  (which corresponds to the linearization method) and in updating the  $H_k$  one needs to see that the positive definiteness has not been violated.

Moreover, the convergence of Newton's method is essentially local. We can circumvent this shortcoming by combining the linearization method with Newton's method. These combinations of the methods have the form (7), where  $H_k$  is some "intermediate" matrix between the  $L''_{xx}(x^k, y^k)$  and  $(1/\gamma)I$ , say, their sum, as in the Levenberg-Marquardt method (see (16) in Section 3.1) for unconstrained minimization. However, even the linearization method has no global convergence, and therefore the combining technique does not guarantee a success each time.

### 9.4.3 Other Methods

Since the most powerful methods for solving nonlinear programming problems (such as Newton's method) are indeed imperfect, it is natural to turn to simpler, more robust methods. Next, we discuss briefly the most important ones.

1. **Penalty Function Methods.** One of the simplest variants of the penalty function method is the one in which all the constraints are taken into account by means of quadratic penalties. In other words, one solves the sequence of auxiliary unconstrained minimization problems

$$\begin{aligned} \min_{x \in \mathbb{R}^n} f_k(x), \\ f_k(x) = f(x) + \frac{1}{2} K_k \left[ \sum_{i=1}^r g_i(x)_+^2 + \sum_{i=r+1}^m g_i(x)^2 \right], \quad K_k \rightarrow \infty. \end{aligned} \quad (10)$$

**THEOREM 4.** Let  $x^*$  be a locally unique solution to problem (1) and let the functions  $f(x)$ ,  $g_i(x)$  be continuously differentiable in a neighborhood of  $x^*$ . Then for sufficiently large  $K_k$  we can find a local minimum point  $x^k$  of  $f_k(x)$  in a neighborhood of  $x^*$ , and  $x^k \rightarrow x^*$ .  $\square$

The following are some specific characteristics of method (10) and of the respective result on its convergence: (1) the convergence has been proven under very weak assumptions, viz. neither the regularity condition nor, *a fortiori*, the nonsingularity condition of the minimum is required; the assumptions concerning the smoothness are not too strong, either. The condition for uniqueness is somewhat restrictive; however even without it one can derive certain convergence results. The fact that the penalty function method gives an approximation for both the solution and the Lagrange multipliers is of importance, too. In fact, if the minimum is nonsingular, then  $K_k g_i(x^k) \rightarrow y_i^*$ ,  $i = 1, \dots, r$ ,  $K_k g_i(x^k) \rightarrow y_i^*$ ,  $i = r+1, \dots, m$ ,  $y_i^*$  being the Lagrange multipliers (cf. Theorem 7 in Section 8.2). Hence (10) can be used together with any more powerful minimization method which requires a good initial approximation in the primal and dual variables.

14

The penalty function method for problem (1) has, of course, the same drawback as the other methods of this group: for large values of the penalty coefficient the unconstrained minimization problem becomes ill-posed, and thus the auxiliary problems are more difficult to solve. Moreover, these problems are generally multimodal (Theorem 4 only asserts that among the local minima of  $f_k(x)$  there is one close to  $x^*$ ).

What has been said so far about the specific method (10) is equally true of the other similar approaches described in Section 9.3.

**2. Dual Methods.** It is not hard to combine the methods described for particular problems (either with equalities or inequalities) as to make them suitable for the general problem (1). We leave it to the interested reader as an exercise. However, these methods can hardly be classified as “robust and reliable” methods for the reasons discussed above.

**3. Reduced Gradient Method.** A rather general approach to solving problem (1) consists in the following. Suppose the approximation  $x^k$  has been obtained. Then one can single out the  $\varepsilon$ -active constraints (i.e., the equalities and those inequalities for which  $|g_i(x^k)| \leq \varepsilon$ ,  $\varepsilon > 0$ ), drop the remaining constraints, and apply, for the problem with equalities, one iteration of either method described in Section 8.2. (Earlier, in the linearization method and in Newton's method we proceeded somewhat differently: the active constraint inequalities were included into the auxiliary problems as inequalities.) It is frequently convenient to use the reduced gradient method in this role (Section 8.2). This, of course, leads to numerous complex problems (possible divergence for a poor initial approximation, the choice of rules for governing  $\varepsilon$  and the accuracy of solution of the auxiliary problems, etc.). Nevertheless, the reduced gradient method is one of the most commonly used methods in software for solving the general mathematical programming problem.

## CHAPTER 10

### LINEAR AND QUADRATIC PROGRAMMING

In the preceding chapters we have already dealt with linear programming or quadratic programming problems while solving more general problems of convex or nonlinear programming, as well as in unconstrained minimization of nonsmooth functions. Problems of linear and quadratic programming are of great interest of their own since they play a major role in economics, engineering, and other areas of application. The mathematical theory of such problems is simple, elegant and complete. The general solution methods with respect to problems of linear and quadratic programming is graphic and most advantageous.

#### 10.1 LINEAR PROGRAMMING (THEORY)

##### 10.1.1 Types of Problems

In the *linear programming problem*, the objective function and the constraints are linear:

$$\begin{aligned} \min (c, x) , \quad & x \in \mathbf{R}^n , \\ (a^i, x) \leq b_i , \quad & i = 1, \dots, m , \end{aligned}$$

or in matrix form:

$$\begin{aligned} \min (c, x) , \quad & x \in \mathbf{R}^n , \\ Ax \leq b , \quad & \end{aligned} \tag{1}$$

where  $A$  is an  $m \times n$  matrix with rows  $a^i$ ;  $c, b$  are fixed vectors in  $\mathbf{R}^n, \mathbf{R}^m$ .

If the condition that the variables be nonnegative is among the constraints, it is more appropriate to keep this condition separate from the matrix  $A$ :

$$\begin{aligned} \min (c, x) , & \quad x \in \mathbf{R}^n , \\ Ax \leq b , & \quad x \geq 0 . \end{aligned} \tag{2}$$

Constraints in the form of equalities are admissible; for example, the problem may have the form:

$$\begin{aligned} \min (c, x) , & \quad x \in \mathbf{R}^n , \\ Ax = b , & \quad x \geq 0 . \end{aligned} \tag{3}$$

Finally, the most general problems are those with *mixed constraints*:

$$\begin{aligned} \min (c, x) , & \quad x \in \mathbf{R}^n , \\ A_1x = b^1 , & \quad A_2x \leq b^2 , \quad x \geq 0 . \end{aligned} \tag{4}$$

Problems of each type can be transformed in the following way. An equality constraint

$$(a^i, x) = b_i \tag{5}$$

is equivalent to the two inequalities

$$(a^i, x) \leq b_i , \quad -(a^i, x) \leq -b_i . \tag{6}$$

The inequality

$$(a^i, x) \leq b_i \tag{7}$$

can be an equality and the nonnegativity condition of the slack variable  $z_i$ : V made

$$(a^i, x) + z_i = b_i , \quad z_i \geq 0 . \tag{8}$$

Also, in a system of equalities one may express some of the variables in terms of the other, and eliminate the former. Thus, if the constraints have the form

$$Ax = b , \quad x \geq 0 , \tag{9}$$

where  $A$  is an  $m \times n$  matrix,  $m < n$ , then by representing it in the form  $A = (A_1 \mid A_2)$ , where  $A_1$  is an  $m \times m$  matrix,  $A_2$  is an  $m \times (n-m)$  matrix, and dividing the variables  $x$  into two groups  $x = \{u, v\}$ ,  $u \in \mathbf{R}^m$ ,  $v \in \mathbf{R}^{n-m}$ , we can write (9) in the form

$$A_1u + A_2v = b , \quad u \geq 0, \quad v \geq 0 .$$

Eliminating  $u$  from the system of equalities (which is possible if  $A_1$  is nonsingular), we obtain constraints in the form of inequalities for  $v$ :

$$A_1^{-1} A_2 v \leq A_1^{-1} b, \quad v \geq 0. \quad (10)$$

Finally, one can always do so as to make the objective function depend on one variable only. Indeed, in the problem

$$\min_{\mathcal{N}} \langle c, x \rangle, \quad x \in Q, \quad (11)$$

where  $Q$  is given by linear constraints, it is possible to introduce the additional variable  $t = (c, x) \in \mathbf{R}^1$  and obtain thus the problem

$$\begin{aligned} & \min t, \\ & (c, x) - t = 0, \quad x \in Q. \end{aligned} \quad (12)$$

The described transformations make it possible to extend any result or any method of solution to problems written in arbitrary form.

Let us comment on the geometric interpretation of the linear programming problem. The linear constraints define a polyhedral set (not necessarily bounded) in the space of variables, and the problem consists in minimizing a linear function on this set (Fig. 41(a)). For (3) the admissible set is formed by the intersection of a linear manifold with the nonnegative orthant (Fig. 41(b)). Finally, (12) is a problem of finding the “lowest point” of a polyhedral set (Fig. 41(c)).

### 10.1.2 Structure of Polyhedral Sets

We need here some auxiliary material pertaining to the theory of linear inequalities (in the geometric terms, to the theory of polyhedral sets). Until now we defined a polyhedral set  $Q$  to be the set of solutions of a system of linear inequalities:

$$Q = \{x \in \mathbf{R}^n : (a^i, x) \leq b_i, i = 1, \dots, m\}, \quad (13)$$

which define the faces of  $Q$  (Fig. 41(a)). In what follows we shall prove that a polyhedral set permits another representation as well, viz. as the convex hull of its vertices (if  $Q$  is not bounded, then  $Q$  is the convex hull of its vertices and of a finite number of rays) (Fig. 42(a)).

A point  $x \in Q$  is called an *extreme point* of the set  $Q \subset \mathbf{R}^n$  if it is not an interior point of any segment lying in  $Q$  (Fig. 42(b)). It is not hard to describe the extreme points of a polyhedral set in algebraic terms.

**LEMMA 1.** The extreme points are exactly the points of the set (13) for which there are  $n$  linearly independent active constraints.

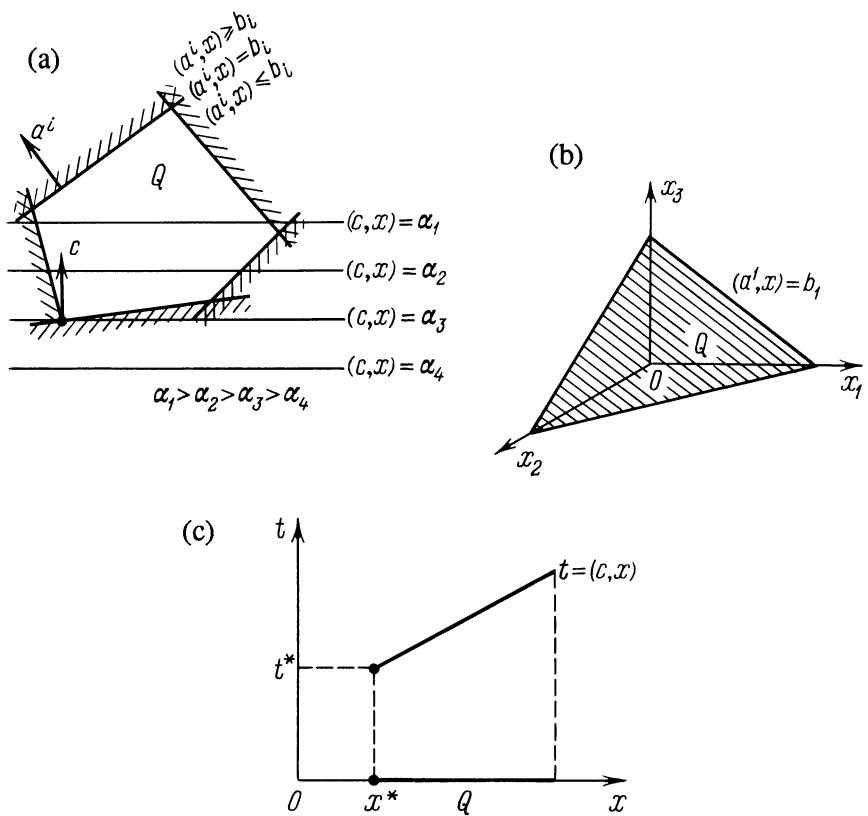


Fig. 41 A linear programming problem: (a) in the form of (1); (b) in the form of (3); (c) in the form of (12).

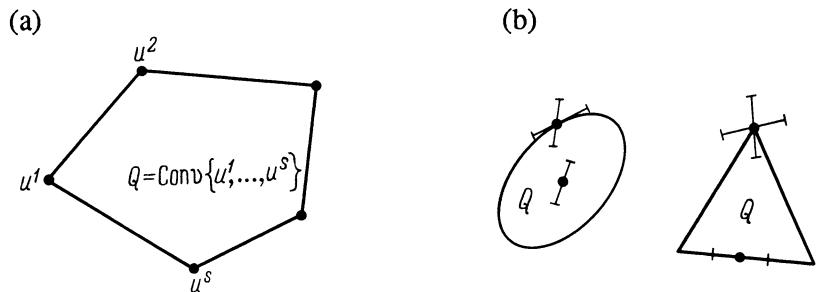


Fig. 42 ~~Re~~ The representation theorem: (a) the convex hull of a finite number of points; (b) extreme points of a set.

Л Т

**PROOF.** Let  $x^0 \in Q$ , i.e.,  $(a^i, x^0) \leq b_i$ ,  $i = 1, \dots, m$ ,  $I_0 = \{i: (a^i, x^0) = b_i\}$  is the set of active constraints. If among the vectors  $a^i$ ,  $i \in I_0$ , fewer than  $n$  vectors are linearly independent, then the system of homogeneous equations  $(a^i, s) = 0$ ,  $i \in I_0$ , has a nonzero solution  $s^0$ . Then the vectors  $x^1 = x^0 + \gamma s^0$ ,  $x^2 = x^0 - \gamma s^0$  for sufficiently small  $\gamma > 0$  also lie in  $Q$ , and therefore the point  $x^0 = (x^1 + x^2)/2$  cannot be an extreme point.

Conversely, suppose that among the  $a^i$ ,  $i \in I_0$ , there are  $n$  linearly independent ones. Suppose that  $x^0 = (x^1 + x^2)/2$ , where  $x^1, x^2 \in Q$ . Then for  $i \in I_0$ , one has

$$b_i = (a^i, x^0) = [(a^i, x^1) + (a^i, x^2)]/2 \leq (b_i + b_i)/2 = b_i.$$

Hence  $(a^i, x^1) = (a^i, x^2) = b_i$ . But the system of equations  $(a^i, x) = b_i$ ,  $i \in I_0$ , cannot have two distinct solutions ( $n$  being its rank). Hence  $x^1 = x^2$ , i.e.,  $x^0$  is an extreme point.  $\square$

**COROLLARY.** The number of extreme points of a polyhedral set is finite.  $\square$

Extreme points of a polyhedral set are called *vertices*. Lemma 1 shows that this terminology is consistent with the geometric definition of a vertex as a 0-dimensional face.

A polyhedral set may or may not have extreme points (e.g., if it is a subspace, see Exercise 1). However, for a bounded  $Q$  extreme points exist and, *a fortiori*, they “generate” the whole set  $Q$ .

**LEMMA 2.** A nonempty bounded polyhedral set is the convex hull of its vertices.

**PROOF.** Let  $x^0 \in Q$ ,  $I_0 = \{i: (a^i, x^0) = b_i\}$ . If  $x^0$  is not an extreme point, we can find  $z^1, z^2 \in Q$ ,  $z^1 \neq z^2$  such that  $x^0 = (z^1 + z^2)/2$ . Consider the line  $x^0 + \lambda(z^2 - z^1)$ . By the boundedness of  $Q$ , it cannot lie entirely inside  $Q$ , therefore at some point  $x^1 = x^0 + \lambda_1(z^2 - z^1)$  it goes out of  $Q$ . Here  $I_1 = \{i: (a^i, x^1) = b_i\} \supset I_0$ , since if  $i \in I_0$ , then  $(a^i, (z^1 + z^2)/2) = b_i$ ,  $(a^i, z^1) \leq b_i$ ,  $(a^i, z^2) \leq b_i$ , which implies  $(a^i, z^1) = (a^i, z^2) = b_i$ , and therefore  $(a^i, x^1) = (a^i, x^0) + \lambda_1[(a^i, z^2) - (a^i, z^1)] = b_i$ . Furthermore, the inclusion  $I_1 \supset I_0$  is strict since at least one active constraint is added at the point  $x^1$ . From this point  $x^1$  we repeat the procedure and arrive at a point  $x^2$ ,  $I_2 \supset I_1$ , and so on. Since the number of active constraints cannot be increased infinitely, we do arrive at an extreme point. Thus extreme points exist; in this case, as the construction implies, for any point  $x^0$  we can find an extreme point  $x^*$ ,  $I_0 \subset I^* = \{i: (a^i, x^*) = b_i\}$ .

Let us show now that any point  $x^0 \in Q$  can be represented in the form  $x^0 = \sum_{i=1}^s \lambda_i u^i$ ,  $\lambda_i \geq 0$ ,  $\sum_{i=1}^s \lambda_i = 1$ , where  $u^1, \dots, u^s$  are extreme points. If

$x^0$  is not an extreme point, we can find a vertex  $u^1 \neq x^0$  (such vertex exists according to what we have proved), take the segment joining  $u^1$  and  $x^0$  and continue it in the direction of  $x^0$  as far as possible (i.e., we find  $x^1 = u^1 + \lambda^*(x^0 - u^1)$ ,  $\lambda^* = \max \{\lambda: u_1 + \lambda(x^0 - u^1) \in Q\} < \infty$ ). At  $x^1$  the set of active constraints  $I_1 \neq \emptyset$ . We find the extreme point  $u^2$  for which  $I_1 \subset \{i: (a^i, u^2) = b_i\}$ . Next, we join  $u^2$  with  $x^1$  and extend this segment toward  $x^1$  as far as possible; we obtain  $x^2$ ,  $I_2 \supset I_1$ ,  $I_2 \neq I_1$ . Continuing this process, we find points  $x^3, x^4, \dots, x^s$  and extreme points  $u^3, \dots, u^s$ . In this case for some  $s$  we have  $x^s = u^s$  (since the set of active constraints is expanding). It is easy to prove by induction that  $x^0$  is a convex combination of  $u^1, \dots, u^s$ .  $\square$

Turning to unbounded polyhedral sets, we start out with the simplest case, viz. a *polyhedral cone*

$$Q = \{x \in \mathbf{R}^n: (a^i, x) \leq 0, i = 1, \dots, m\}, \quad (14)$$

and assume that  $A$  is a matrix with rows  $a^1, \dots, a^m$ .

**LEMMA 3.** The cone  $Q$  in (14) can be represented as the convex hull of a finite number of rays and lines:

$$Q = \left\{ x: x = \sum_{i=1}^s \lambda_i u^i + \sum_{i=1}^r \gamma_i w^i, \lambda_i \geq 0, \lambda_i, \gamma_i \in \mathbf{R}^1 \right\}, \quad (15)$$

$$u^i, w^i \in Q. \quad \square$$

Let us introduce some notions regarding an arbitrary cone  $K$ . A ray  $\{\lambda v, \lambda \geq 0\}$ ,  $v \in K$ , is said to be an extreme ray for  $K$  if it cannot be represented as a half-sum of two distinct rays in  $K$ . Extreme rays (they are called *directors*) play the same role for cones as extreme points do for bounded sets. A cone  $K$  is said to be *sharp* if it contains no lines, i.e., the fact that  $v, -v \in K$  implies  $v = 0$ . Using this terminology, we derive from Lemma 3 an analog of Lemma 2 for cones.

**LEMMA 4.** A sharp polyhedral cone is the convex hull of its extreme rays, the number of which is finite.  $\square$

We can now obtain the main result of this section.

**THEOREM 1** (on representation of a polyhedral set). Let

$$Q = \{x \in \mathbf{R}^n: Ax \leq b\} \neq \emptyset. \quad (16)$$

Then  $Q$  is representable as the convex hull of a finite number of points, rays and lines:

$$Q = \left\{ x: x = \sum_{i=1}^s \lambda_i u^i + \sum_{i=1}^r \mu_i v^i + \sum_{i=1}^p \gamma_i w^i, \lambda_i \geq 0, \sum_{i=1}^s \lambda_i = 1, \mu_i \geq 0, \gamma_i, \lambda_i, \mu_i \in \mathbf{R}^1 \right\}, \quad (17)$$

where  $Au^i \leq 0$ ,  $i = 1, \dots, r$ ;  $Aw^i = 0$ ,  $i = 1, \dots, p$ ;  $(w^i, w^j) = 0$ ,  $i \neq j$ . If  $Q$  contains no lines ( $p = 0$ ), then for  $u^i$  one can take the vertices of  $Q$ , and as  $v^i$  the extreme rays of  $K = \{x: Ax \leq 0\}$ .

The proof is a combination of the arguments made above. First we find  $w^1, \dots, w^p$  such that  $Aw^i = 0$ ,  $i = 1, \dots, p$ ;  $(w^i, w^j) = 0$ ,  $i \neq j$  (if there are such). Next we consider the polyhedral set

$$Q_1 = \{x: A(x^0 + x) \leq b, (x, w^i) = 0, i = 1, \dots, p\},$$

where  $x^0$  is any point in  $Q$ : then  $0 \in Q_1$ , i.e.,  $Q_1 \neq \emptyset$ , and we show that  $Q_1$  is representable as the convex hull of its extreme points and the directors of the cone  $K = \{x: Ax \leq 0, (x, w^i) = 0, i = 1, \dots, p\}$ , which is done just as in the proof of Lemma 2 but taking into account that  $Q_1$  may contain rays. The rays are represented as the convex hull of the directors of  $K$  by means of Lemma 3.  $\square$

A vector  $z \in \mathbf{R}^n$  such that the ray  $x + \lambda z$ ,  $\lambda \geq 0$ ,  $x \in Q$ , lies entirely in  $Q$  is called a *direction of recession* of the convex set  $Q$ . It follows from Theorem 1 that for a polyhedral set all directions of recession are generated by a finite set of vectors.

The result dual to Theorem 1 is also valid: the convex hull of a finite number of points, rays and lines is a polyhedral set, i.e., representable in the form (13).

## Exercises

1. Prove the following assertions:

(a) every extreme point of  $Q$  is a boundary point of  $Q$ ; in particular, an open set has no extreme points;

(b) a subspace has no extreme points;

(c) every point of the sphere  $\{x: \|x\| = 1\}$  is an extreme point for the ball  $Q = \{x: \|x\| \leq 1\}$ .

2. Use Theorem 1 to show that for  $Q = \{x: Ax \leq b\} \neq \emptyset$  the following properties are equivalent:

(a)  $Q$  is bounded;

(b) the inequalities  $Ax \leq 0$  have only the null solution;

(c) the equations  $A^T y = c$  have a solution  $y \geq 0$  for any  $c$  (use Farkas' lemma).

### 10.1.3 Extremum Conditions

Theoretically, Theorem 1 enables one to find a solution of the linear programming problem and to investigate it completely.

**THEOREM 2.** The solution set  $X^*$  of the problem

$$\begin{aligned} \min (c, x), & \quad x \in \mathbf{R}^n, \\ x \in Q = \{x: Ax \leq b\}, & \quad b \in \mathbf{R}^m, \end{aligned} \quad (18)$$

is nonempty iff  $Q \neq \emptyset$  and  $(c, v^i) \geq 0$ ,  $i = 1, \dots, r$ ;  $(c, w^i) = 0$ ,  $i = 1, \dots, p$ , and is given by

$$\begin{aligned} X^* = \left\{ x^*: x^* = \sum_{i \in I_1} \lambda_i u^i + \sum_{i \in I_2} \mu_i v^i + \sum_{i=1}^p \gamma_i w^i, \lambda_i \geq 0, \right. \\ \left. \sum_{i \in I_1} \lambda_i = 1, \mu_i \geq 0, \gamma_i \in \mathbf{R}^1 \right\}, \end{aligned} \quad (19)$$

where

$$I_1 = \{i: (u^i, c) = f^*\}, \quad I_2 = \{i: (c, v^i) = 0\}, \quad f^* = (c, x^*), \quad x^* \in X^*,$$

and  $u^i, v^i, w^i$  are the vectors in the statement of Theorem 1.

Indeed, by (17) for any  $x \in Q$  one has

$$(c, x) = \sum_{i=1}^s \lambda_i (c, u^i) + \sum_{i=1}^r \mu_i (c, v^i) + \sum_{i=1}^p \gamma_i (c, w^i). \quad (20)$$

The minimum of this expression for  $\lambda_i \geq 0$ ,  $\sum_{i=1}^s \lambda_i = 1$ ,  $\mu_i \geq 0$ ,  $\gamma_i \in \mathbf{R}^1$ , obviously obtains only when  $(c, v^i) \geq 0$ ,  $i = 1, \dots, r$ ,  $(c, w^i) = 0$ ,  $i = 1, \dots, p$ , and the solution is given by  $\lambda_i^*, \mu_i^*, \gamma_i^*$ , such that  $\lambda_i^* = 0$  for  $i \notin I_1$ ,  $\lambda_i^* \geq 0$ ,  $\sum_{i \in I_1} \lambda_i^* = 1$ ,  $\mu_i^* \geq 0$  for  $i \in I_2$ ,  $\mu_i^* = 0$  for  $i \notin I_2$ .  $\square$

**COROLLARY.** If the admissible set contains no lines and, furthermore, a solution of problem (18) exists, we can find a vertex of  $Q$  among the solutions of (18). If the solution is unique, it is at a vertex.  $\square$

It goes without saying that this result cannot be viewed as a constructive way of finding a solution, since the number of vertices is great even for small-scale problems and it is not possible to go through the entire class of vertices. Thus, Theorem 2 does not make the problem of finding solutions to linear programming problems trivial. Hence it makes sense to analyze the problem in the ordinary way, which involves extremum conditions. Problem (18) is a very special case of the convex programming

problem (Section 9.1): the objective function and the constraints are defined for all  $x \in \mathbf{R}^n$ , both convex and differentiable, and the constraints have special form, making it possible to dispense with Slater's condition. Using Theorem 2 and Theorem 3 of Section 9.1, we arrive at the following extremum conditions.

**THEOREM 3.** For an admissible point  $x^*$  to be a solution of problem (18) it is necessary and sufficient that there exist Lagrange multipliers  $y^* \in \mathbf{R}^m$  such that

$$y^* \geq 0, \quad (y^*, Ax^* - b) = 0, \quad c + A^T y^* = 0. \quad \square \quad (21)$$

**THEOREM 4.** For  $x^* \in \mathbf{R}^n$  to be a solution of problem (18) it is necessary and sufficient that there exist  $y^* \in \mathbf{R}^m$ ,  $y^* \geq 0$ , such that

$$L(x, y^*) \geq L(x^*, y^*) \geq L(x^*, y) \quad \forall x \in \mathbf{R}^n, \quad \forall y \in \mathbf{R}_+^m, \quad (22)$$

where

$$L(x, y) = (c, x) + (y, Ax - b). \quad \square \quad (23)$$

Let us write the dual problem of (18) and formulate the duality theorem. Since

$$\psi(y) = \inf_{x \in \mathbf{R}^n} L(x, y) = \inf_{x \in \mathbf{R}^n} [(c + A^T y, x) - (b, y)],$$

then

$$\psi(y) = \begin{cases} -\infty & \text{if } c + A^T y \neq 0, \\ -(b, y) & \text{if } c + A^T y = 0. \end{cases}$$

Writing the dual problem in the form (49) of Section 9.1, we obtain

$$\min (b, y), \quad c + A^T y = 0, \quad y \geq 0. \quad (24)$$

Thus, the dual problem of (18) is a linear programming problem.

**THEOREM 5** (duality theorem). Solutions  $x^*$ ,  $y^*$  of the dual problems (18) and (24) exist, or do not exist, simultaneously, and  $y^*$  (respectively  $x^*$ ) are the Lagrange multipliers for (18) (respectively for (24)). For any admissible  $x$ ,  $y$  we have the inequality

$$(c, x) + (b, y) \geq 0, \quad (25)$$

with equality iff  $x$ ,  $y$  are solutions to (18) and (24).  $\square$

The reader is advised to prove this theorem directly using Theorems 3 and 4, rather than invoking the general duality theorem of Section 9.1. It is convenient in this case to use the extremum conditions given below for a problem of the form (24) (see Theorem 6).

The results of Theorems 3, 4 and 5 are related to the problem of the form (18). If the initial problem is given in a different form, then, using the transformations described in Subsection 10.1.1, we can reduce the problem to that of the form (18), write the extremum conditions and, next, return to the original variables. If we do all this for the problem

$$\begin{aligned} \min (c, x), \quad & x \in \mathbf{R}^n, \\ Ax = b, \quad & b \in \mathbf{R}^m, \quad x \geq 0 \end{aligned} \tag{26}$$

(called the *canonical* form of a linear programming problem), we obtain the following results.

**THEOREM 6.** The point  $x^* \in \mathbf{R}^n$  is a solution of the linear programming problem (26) if and only if we can find  $y^* \in \mathbf{R}^m$  such that either of the following conditions holds:

(a)

$$Ax^* = b, \quad x^* \geq 0, \quad A^T y^* \leq c, \quad (A^T y^* - c, x^*) = 0, \tag{27}$$

(b)

$$x^* \geq 0, \quad L(x, y^*) \geq L(x^*, y^*) \geq L(x^*, y), \tag{28}$$

$$\forall x \in \mathbf{R}_+^n, \quad \forall y \in \mathbf{R}^m. \quad \square$$

The dual problem of (26) is

$$\begin{aligned} \max (b, y), \\ A^T y \leq c. \end{aligned} \tag{29}$$

**THEOREM 7** (duality theorem). Solutions  $x^*, y^*$  of problems (26) and (29) exist, or do not exist, simultaneously, and in this case  $y^*$  (respectively  $x^*$ ) are the Lagrange multipliers for (26) (respectively for (29)). For any admissible  $x, y$  we have

$$(c, x) \geq (b, y), \tag{30}$$

with equality iff  $x, y$  are solutions of (26) and (29).  $\square$

Now that we know how to write the dual problem for different formulations of the original problem, we can establish the following result which justifies the term “duality.”

**THEOREM 8.** The primal problem is dual to the dual problem (24). The same is true for problems (26) and (29).  $\square$

Therefore, we can speak of a pair of dual problems without identifying the primal problem. We note also that the method for constructing the dual problem, as described in Section 9.1, does not allow us to make the same statement concerning the general convex programming problem, since we do not know how to construct the dual problem of a dual problem.

A pair of dual problems can be written in the symmetric form:

$$\begin{array}{ll} \min(c, x), & \max(b, y), \\ Ax \geq b, & A^T y \leq c, \\ x \geq 0; & y \leq 0, \end{array} \quad (31)$$

where  $(c, x) \geq (b, y)$  for admissible  $x, y$ , and we have equality only for the solutions.

### Exercise

3. Use Theorem 2 to show that if  $Q$  contains lines, a solution of (18) cannot be bounded. Derive necessary and sufficient conditions for a solution to be bounded.

#### 10.1.4 Existence, Uniqueness and Stability of a Solution

As noted earlier, Theorem 2 resolves the question of the existence and uniqueness, but it is not constructive. Theorem 2 can be used to obtain more easily verifiable conditions for a solution to exist and be unique. This is what we are interested in now.

**THEOREM 9.** Suppose that in problem (18) the objective function is bounded below on a nonempty admissible set:

$$Q = \{x: Ax \leq b\} \neq \emptyset, \quad \inf_{x \in Q} (c, x) > -\infty.$$

Then a solution of problem (18) exists.

**PROOF.** The function  $(c, x)$  on  $Q$  is given by (20). However, (20) is bounded below for any  $\mu_i \geq 0, \gamma_i \in \mathbf{R}^1$  only if  $(c, v^i) \geq 0, i = 1, \dots, r; (c, w^i) = 0, i = 1, \dots, p$ , with the minimum of  $(c, x)$  obtained on  $Q$  (Theorem 2).  $\square$

We emphasize the fact that in this theorem we do not assume that sets of the form  $\{x \in Q: (c, x) \leq \alpha\}$  are bounded.

As a corollary of Theorem 9 we derive that the problem of unconstrained minimization of each of the following piecewise linear functions:

$$\begin{aligned} f(x) &= \sum_{i=1}^m ((a^i, x) - b_i)_+ , & f(x) &= \sum_{i=1}^m |(a^i, x) - b_i| , \\ f(x) &= \max_{1 \leq i \leq m} ((a^i, x) - b_i)_+ , & f(x) &= \max_{1 \leq i \leq m} |(a^i, x) - b_i| \end{aligned} \quad (32)$$

has a solution. Indeed, these problems reduce to linear programming problems in which the objective function is nonnegative.

Another useful corollary of Theorem 9 is a refined duality theorem.

**THEOREM 10.** If both dual linear programming problems have admissible points, these problems have solutions.

Indeed, this result follows immediately from relation (25) (for problems (18) and (24)) and Theorem 9.  $\square$

It is very easy to see whether a given solution is unique.

**THEOREM 11.** Let  $x^*$  be a solution to problem (1) and let  $I^* = \{i: (a^i, x^*) = b_i\}$  be the set of active constraints. Then, if among the vectors  $a^i$ ,  $i \in I^*$ , we cannot find  $n$  linearly independent ones, then  $x^*$  is not unique, whereas if there are  $n$  such vectors and the corresponding  $y_i^*$  are positive, then the solution is unique.

**PROOF.** By a corollary of Theorem 2 a unique minimum is attained at a vertex, which together with Lemma 1 gives the first assertion of the theorem. Next, let  $a^i$ ,  $i \in I \subset I^*$ , be  $n$  linearly independent vectors with indices in  $I^*$ . Then the equation  $(a^i, x) = b_i$ ,  $i \in I$ , has a unique solution  $x^*$ . Hence for any  $x \in Q$ ,  $x \neq x^*$ , we can find a  $j \in I$  such that  $(a^j, x) < b_j$ . Then by Theorem 3,

$$\begin{aligned} (c, x) &= -(A^T y^*, x) = -\sum_{i \in I^*} y_i^*(a^i, x) > -\sum_{i \in I} y_i^* b_i = \\ &= -\sum_{i \in I} y_i^*(a^i, x^*) = -(A^T y^*, x^*) = (c, x^*) . \quad \square \end{aligned}$$

With regard to stability of solution, it turns out that for linear programming problems one has a relation which generalizes the sharp minimum condition (11) of Section 7.1.

**LEMMA 5.** Let  $X^*$  be the set of solutions to problem (18),  $X^* \neq \emptyset$ . Then we can find an  $\alpha > 0$  such that

$$(c, x) - f^* \geq \alpha \rho(x, X^*) \quad (33)$$

for all admissible  $x$ . In this case

$$f^* = (c, x^*), \quad x^* \in X^*, \quad \rho(x, X^*) = \min_{x^* \in X^*} \|x - x^*\|.$$

We prove the case where  $Q$  is bounded.

We have

$$\begin{aligned} Q &= \left\{ x: x = \sum_{i=1}^s \lambda_i u^i, \lambda_i \geq 0, \sum_{i=1}^s \lambda_i = 1 \right\}, \\ X^* &= \left\{ x: x = \sum_{i \in I_1} \lambda_i^* u^i, \lambda_i^* \geq 0, \sum_{i \in I_1} \lambda_i^* = 1 \right\}, \end{aligned}$$

where  $I_1 = \{i: (c, u^i) = f^*\}$  (see Theorems 1 and 2). Introduce

$$\alpha = \min_{i \in I_1} \frac{(c, u^i) - f^*}{\rho(u^i, X^*)}.$$

Let us show that this is the required  $\alpha$ . Indeed,  $\alpha > 0$  since  $(c, u^i) - f^* > 0$  for all  $i \notin I_1$ . Let  $x \in Q$  be arbitrary,  $x = \sum_{i=1}^s \lambda_i u^i$ ,  $\lambda_i \geq 0$ ,  $\sum_{i=1}^s \lambda_i = 1$ . Applying Jensen's inequality (Lemma 1 of Section 1.1) to the convex function  $\rho(x, X^*)$  (see Exercise 2 in Section 5.1), we obtain

$$\begin{aligned} \rho(x, X^*) &= \rho\left(\sum_{i=1}^s \lambda_i u^i, X^*\right) \leq \sum_{i=1}^s \lambda_i \rho(u^i, X^*) = \sum_{i \in I_1} \lambda_i \rho(u^i, X^*) \\ &\leq \alpha^{-1} \sum_{i \in I_1} \lambda_i ((c, u^i) - f^*) = \alpha^{-1} \sum_{i=1}^s \lambda_i ((c, u^i) - f^*) \\ &= \alpha^{-1} (c, x) - f^*. \quad \square \end{aligned}$$

Lemma 5 implies in particular that a unique solution of a linear programming problem satisfies the sharp minimum condition:

$$(c, x) - f^* = \alpha \|x - x^*\| \quad \forall x \in Q. \quad (34)$$

**LEMMA 6 (Hoffman).** Let

$$Q = \{x \in \mathbf{R}^n: (a^i, x) \leq b_i, i = 1, \dots, m\} \neq \emptyset.$$

Then we can find an  $\alpha > 0$  such that

$$\sum_{i=1}^m ((a^i, x) - b_i)_+ \geq \alpha \rho(x, Q) \quad \forall x \in \mathbf{R}^n. \quad (35)$$

**PROOF.** Consider the linear programming problem

$$\begin{aligned} \min & \sum_{i=1}^m t_i, \\ (a^i, x) - b_i &= t_i - z_i, \quad i = 1, \dots, m, \\ t_i \geq 0, \quad z_i \geq 0, & \quad i = 1, \dots, m, \end{aligned} \tag{36}$$

where  $t_i, z_i$  are supplementary variables. For any  $x \in \mathbb{R}^n$  the vector  $\{x, t, z\}$ , where  $t_i = ((a^i, x) - b_i)_+$ ,  $z_i = (b_i - (a^i, x))_+$ ,  $i = 1, \dots, m$ , is an admissible point in this problem, and the solution to (36) has the form  $\{x, 0, (b - Ax)_+\}$ , where  $x \in Q$ . Applying Lemma 5 to this problem yields (35).  $\square$

Hoffman's lemma states that if at some point the residual by which the linear equalities are violated is small, then this point is close to the set of admissible points. Note that a similar result for problems with convex inequalities (Lemma 11 of Section 9.1) was derived under the assumption that Slater's condition holds and  $Q$  is bounded.

For the linear programming problem Lemma 6 yields the following assertion on stability.

**LEMMA 7.** Let  $X^*$  be the nonempty set of solutions of problem (1),  $f^* = (c, x^*)$ ,  $x^* \in X^*$ . Then we can find an  $\alpha > 0$  such that for any  $x$ ,

$$\alpha\rho(x, X^*) \leq \sum_{i=1}^m ((a^i, x) - b_i)_+ + ((c, x) - f^*)_+. \tag{37}$$

To prove this lemma, it suffices to write  $X^*$  in the form  $X^* = \{x: (a^i, x) \leq b_i, i = 1, \dots, m, (c, x) \leq f^*\}$  and use Lemma 6.  $\square$

It follows from (37) that every generalized minimizing sequence  $x^k$  (i.e., such that  $(c, x^k) \rightarrow f^*$ ,  $((a^i, x^k) - b_i)_+ \rightarrow 0$ ) converges to the set of solutions of the linear programming problem.

The fact that in a linear programming problem the extended sharp minimum condition (33) is satisfied implies that a solution is invariant toward a small perturbation of the objective function.

**THEOREM 12.** Let the set  $X^*$  of solutions of problem (18) be nonempty, let the function  $g(x)$  be convex, let  $D(g^0) \supset X^*$ ,  $\|\partial g(x)\| \leq L$  for  $x \in X^*$ , and let  $g(x)$  attain a minimum on  $X^*$ :  $X_g^* = \arg \min_{x \in X^*} g(x) \neq \emptyset$ . Then for sufficiently small  $\varepsilon > 0$  the minimum in the problem

$$\min_{x \in Q} [(c, x) + \varepsilon g(x)] \tag{38}$$

obtains on the set  $X_g^*$ .



*PROOF.* Let  $x$  be an arbitrary point in  $Q$ , let  $x^*$  be the projection of  $x$  onto  $X^*$ , and let  $x_g^*$  be a point in  $X^*$ . Then by Lemma 5,

$$(c, x) + \varepsilon g(x) \geq f^* + \alpha \|x - x^*\| + \varepsilon(g(x_g^*) + (\partial g(x_g^*), x - x_g^*)).$$

But

$$(\partial g(x_g^*), x - x_g^*) = (\partial g(x_g^*), x - x^*) + (\partial g(x_g^*), x^* - x_g^*),$$

where

$$(\partial g(x_g^*), x^* - x_g^*) \geq 0$$

by the extremum conditions for  $g(x)$  on  $X^*$ . Hence

$$\begin{aligned} (c, x) + \varepsilon g(x) &\geq f^* + \varepsilon g(x_g^*) + \alpha \|x - x^*\| + \varepsilon(\partial g(x_g^*), x - x^*) \\ &\geq (c, x_g^*) + \varepsilon g(x_g^*) + (\alpha - \varepsilon L)\|x - x^*\| \geq (c, x_g^*) + \varepsilon g(x_g^*) \end{aligned}$$

for  $\varepsilon < \alpha/L$ .  $\square$

As will be shown later, from this important result we can derive the fact that many of the iterative linear programming methods are finite. Applying Theorem 12 to the dual problem, we find that under small perturbations of the constraints, the Lagrange multipliers remain the same (Exercise 4). It follows that Theorem 9 of Section 9.1 may be refined for linear programming problems (Exercise 5).

### Exercises

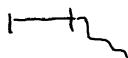
4. Show that if solutions exist for the problems  $\min (c, x)$ ,  $Ax \leq b$  and  $\min (c, x)$ ,  $Ax \leq b + \varepsilon$ , then for sufficiently small  $\varepsilon \in \mathbf{R}^m$  the Lagrange multipliers of the latter problem are also Lagrange multipliers for the former problem.

5. Let the solution  $x^*$  of the problem  $\min (c, x)$ ,  $Ax \leq b$  and let the corresponding Lagrange multipliers  $y^*$  be unique. In this case the function  $\phi(\varepsilon) = \inf_{Ax \leq b + \varepsilon} (c, x)$ ,  $\varepsilon \in \mathbf{R}^m$ , is differentiable at 0 and  $\nabla \phi(0) = y^*$ .

6. Let the sets  $X^* = \underset{Ax \leq b}{\operatorname{Argmin}} (c, x)$ ,  $X_1^* = \underset{x \in X^*}{\operatorname{Argmin}} (c^1, x)$ ,  $X_k^* = \underset{x \in X_{k-1}^*}{\operatorname{Argmin}} (c^k, x)$  be nonempty. Prove that  $x_k^*$  is the solution of the problem  $\min_{Ax \leq b} [(c, x) + \varepsilon(c^1, x) + \dots + \varepsilon^k(c^k, x)]$  for sufficiently small  $\varepsilon > 0$ .

7. Using the preceding result, show that the nonuniqueness of the solution can be removed by a small perturbation, i.e., if a solution of the problem  $\min (c, x)$ ,  $Ax \leq b$ , is nonunique, then we can find  $\varepsilon > 0$  and  $c^1 \in \mathbf{R}^n$  such that the solution of the problem  $\min (c + \varepsilon c^1, x)$ ,  $Ax \leq b$ , is unique.

$\checkmark$  for any  $\varepsilon > 0$



## 10.2 FINITE LINEAR PROGRAMMING METHODS

### 10.2.1 The Simplex Method

As was shown earlier (Corollary of Theorem 2 in Section 10.1) a solution of a linear programming problem is attained at a vertex of a polyhedral set. On the other hand, Lemma 1 of Section 10.1 gives a simple description of the vertices as solutions of systems of linear equations. Since the number of vertices is finite, one can, in principle, find the solution in a finite number of steps (each of which includes solving a system of linear equations, i.e., it requires a finite number of arithmetic operations) by reviewing the vertices.

The simplex method uses essentially the same idea, but not so straightforwardly. (1) The vertices are reviewed in such a way that the values of the objective function decrease monotonically. In this case the vertices at which the values of  $(c, x)$  are greater than the ones already found are discarded. (2) The review includes adjacent (or, neighboring) vertices, and therefore the system of linear equations to be solved at each step differs very little from that to be solved at the preceding step. Furthermore, for solving these systems, special, economical procedures are usually used.

The simplex method is employed for problems written in canonical form:

$$\begin{aligned} \min (c, x), \\ Ax = b, \quad x \geq 0, \end{aligned} \tag{1}$$

where  $x \in \mathbf{R}^n$ ,  $A$  is an  $m \times n$  matrix,  $b \in \mathbf{R}^m$ ,  $c \in \mathbf{R}^n$ . Since in Section 10.1 we dealt basically with a different notation of the admissible set, we reformulate the results obtained therein.

**LEMMA 1.** Let

$$Q = \{x \in \mathbf{R}^n : Ax = b, x \geq 0\} \neq \emptyset, \quad b \in \mathbf{R}^m. \tag{2}$$

Then: (a)  $Q$  is a polyhedral set containing no lines;

(b) the vertices of  $Q$  are the points  $x$  in  $Q$ , for which the vectors  $A^i$ ,  $i \in I$ , are linearly independent ( $A^i$  is the  $i$ th column of  $A$ ,  $I = \{i : x_i \neq 0\}$ );

(c)  $Q$  is representable as the convex hull of its vertices and extreme rays of the cone  $K = \{x \geq 0, Ax = 0\}$ .  $\square$

By Lemma 1, at a vertex of  $Q$  no more than  $m$  components of the vector  $x$  can be positive. We say that  $x$  is a *nonsingular vertex* if the number of its positive components is equal to  $m$ . We shall assume in the sequel that the admissible set in (1) is nonempty and all of its vertices are nonsingular. The simplex method involves specific terminology and it is frequently not uniform. The admissible point is called a *feasible solution*, the vertex is

called a *basic feasible solution*, the solution of the problem is called an *optimal feasible solution*, and the columns of  $A$  that correspond to positive components of a basic feasible solution are called a *basis*. (We shall, however, be using the terms admissible point, vertex, solution, etc.)

Let us proceed to describe the simplex method for solving problem (1). Suppose that at the  $k$ th step we have obtained a point  $x^k$  which is a vertex of (2), and also

$$I_k = \{i: x_i^k > 0\}. \quad (3)$$

By the nonsingularity assumption, the  $I_k$  contains  $m$  components. We decompose the vector  $x \in \mathbb{R}^n$  into two groups  $x = \{u, v\}$ , where  $u \in \mathbb{R}^m$  corresponds to the components in  $I_k$ , and  $v \in \mathbb{R}^{n-m}$  to the components not in  $I_k$ . Then the system  $Ax = b$  can be written

$$A_1 u + A_2 v = b, \quad (4)$$

where  $A_1$  is the  $m \times m$  matrix whose columns are the columns of  $A$  with indices in  $I_k$ ,  $A_2$  is the  $m \times (n-m)$  matrix composed of the remaining columns of  $A$ . By Lemma 1(b) and by the nonsingularity condition, the matrix  $A_1$  has an inverse, and therefore

$$u = A_1^{-1}(b - A_2 v). \quad (5)$$

The objective function becomes

$$(c, x) = (c^1, u) + (c^2, v), \quad (6)$$

where  $c^1 \in \mathbb{R}^m$  is the vector with components  $c_i$ ,  $i \in I_k$ ,  $c^2 \in \mathbb{R}^{n-m}$  is the vector with components  $c_j$ ,  $j \notin I_k$ . Noting (5), the objective function can be expressed solely in terms of  $v$  and thus becomes

$$(c^1, A_1^{-1}(b - A_2 v)) + (c^2, v) = (c^2 - A_2^T(A_1^{-1})^T c^1, v) + (c^1, A_1^{-1} b).$$

Therefore, the initial problem is equivalent to the problem

$$\min (c^2 - A_2^T(A_1^{-1})^T c^1, v), \quad A_1^{-1}(b - A_2 v) \geq 0, \quad v \geq 0, \quad (7)$$

where the vector  $\{u^k, v^k\}$  corresponds to the point  $x^k$ ,  $u^k > 0$ ,  $v^k = 0$ . The point  $v^k = 0$  is admissible for problem (7), and the constraint  $A_1^{-1}(b - A_2 v) \geq 0$  is a strict inequality since  $A_1^{-1}(b - A_2 v^k) = A_1^{-1}b = u^k > 0$ . Hence it can be dropped (since it is inactive) in checking  $v^k$  for optimality. But in the problem

$$\min (d, v), \quad v \geq 0, \quad (8)$$

the minimum obtains at 0 iff  $d \geq 0$ .

Thus, if

$$d = c^2 - A_2^T (A_1^{-1})^T c^1 \geq 0, \quad (9)$$

the point  $v = 0$  is a solution of (8) and at the same time of (7), and hence  $x^k$  is a solution of (1). However, if  $d$  has some negative components (say,  $d_j < 0$ ), then  $v = 0$  is not a solution of (8), and by increasing  $v_j$  we can decrease the value of the objective function in (8). Thus, we make the step

$$v^{k+1} = v^k + \gamma_k e_j, \quad j: d_j < 0, \quad e_j \text{ is the } j\text{th basis vector}. \quad (10)$$

The choice of  $\gamma_k$  is based on the following considerations. As  $\gamma_k$  increases, the objective function decreases; however, the constraint  $A_1^{-1}(b - A_2 v) \geq 0$ , which was an inactive constraint at  $v^k = 0$  may be violated. Hence the  $\gamma_k$  need to be such as to satisfy

$$\gamma_k = \max \{ \gamma \geq 0 : A_1^{-1}(b - \gamma A_2 e_j) \geq 0 \}. \quad (11)$$

If  $\gamma_k = \infty$ , then the problem has no solution since  $\inf_{x \in Q} (c, x) = -\infty$ . But if  $\gamma_k < \infty$ , then we obtain the new point

$$x^{k+1} = \{u^{k+1}, v^{k+1}\},$$

where

$$u^{k+1} = A_1^{-1}(b - \gamma_k A_2 e_j), \quad v^{k+1} = v^k + \gamma_k e_j.$$

This point is again a vertex (it is admissible and has  $m$  positive components since the  $j$ th component has become positive, while one of the components of  $u^{k+1}$  has vanished). Hence, at this point one can repeat the whole procedure once more. Since  $(c, x^{k+1}) < (c, x^k)$  (because  $(c, x^{k+1}) = (c, x^k) + \gamma_k d_j$ ,  $\gamma_k > 0$ ,  $d_j < 0$ ), the objective function is monotonic for the method. Hence the return to any of the "passed" vertices is impossible and since the total number of vertices is finite, the method is finite as well.

In the geometric context, the simplex method is a sequential transition from one vertex of the admissible set to an adjacent one (see Exercise 1), where the objective function has a smaller value. Clearly, such a process is finite.

The idea of elimination of variables that is used in the simplex method was reviewed in Chapter 8 in analyzing problems involving equality constraints. Thus, the second proof of the Lagrange multipliers rule (Section 8.1) was based on a reduction of the initial problem to an unconstrained minimization problem by expressing (implicitly) particular variables in terms of the other variables and using extremum conditions in the unconstrained problem. A similar procedure was employed in the reduced gradient method to construct the optimization method (compare formula (18) in Section 8.2).

with the formulas for the simplex method). Practically the same technique is used in the simplex method, with the only difference that in addition to the equalities  $Ax = b$  there are inequalities  $x \geq 0$ , and each problem resulting from the elimination of variables is a linear programming problem of simple structure (7). Incidentally, the optimality condition (9) in the simplex method enables one to obtain extremum conditions in the initial problem in a different way, although under more stringent assumptions (see Exercise 2).

### Exercises

1. We say that the edge (a one-dimensional face) of a polyhedral set  $Q$  is a segment  $L = [a, b] \subset Q$  such that none of its points is an interior point of another segment in  $Q$  (i.e., if  $x \in L$ ,  $y \in Q$ ,  $z \in Q$  and  $x = (y+z)/2$  are possible only if  $y \in L$ ,  $z \in L$ ). Prove that the endpoints of an edge are vertices (such vertices are called adjacent, or neighboring) and, also, there exists an adjacent vertex (if the number of vertices is greater than 1). Give a description of the vertices in Lemma 1 of Section 10.1 and in Lemma 1(b) in this section. Show that the points  $x^k$  and  $x^{k+1}$  in the simplex method are adjacent vertices.
2. Derive from (9) the extremum conditions in problem (1). Show that they coincide with the conditions stated in Theorem 6 of Section 10.1.

*Hint:* Prove that if  $x^*$  is a solution, then  $(A_1^{-1})^T c^1 = y^*$ .

#### 10.2.2 Implementation of the Simplex Method

Of course the above description of the simplex method is far from complete. In order to go from this description over to a well-defined algorithm, one has to resolve a couple of questions.

1. *The choice of an initial approximation.* A point  $x^0$  which is a vertex of  $Q$  can be found via the following procedure (the *artificial basis method*). Introduce the additional slack variables  $z_i$  which play the role of the residuals in the constraints, and consider the problem of minimizing these variables:

$$\min \sum_{i=1}^m z_i , \quad (12)$$

$$(a^i, x) + z_i = b_i , \quad i = 1, \dots, m, \quad x \geq 0, \quad z \geq 0 .$$

In this problem the required vector is a vector  $\{x, z\}$  of dimension  $n+m$ , while the point  $\{0, b\} \in \mathbb{R}^{n+m}$  is a vertex (by changing the sign of the constraint  $(a^i, x) = b_i$ , if necessary, one can always assume that  $b \geq 0$ ; we suppose that  $b$  has no null components). Hence for (12) it is possible to apply the simplex method with a given initial approximation. As a result, we obtain either the point  $\{x^0, 0\}$ , where  $x^0$  is a vertex in the initial problem, or the point with  $z \neq 0$ ; then there are no admissible vectors (the constraints are contradictory).

One can often avoid this additional operation in solving linear programming problems. For example, suppose that the initial problem is

$$\begin{aligned} \min & (c, x), \\ Ax & \leq b, \quad x \geq 0, \end{aligned} \tag{13}$$

with  $b > 0$ . We put it in the canonical form as in formula (8) in Section 10.1:

$$\begin{aligned} \min & (c, x) \\ Ax + z & = b, \quad x \geq 0, z \geq 0. \end{aligned} \tag{14}$$

Then the point  $\{0, b\}$  is a vertex, so that the artificial basis method is superfluous.

2. *Numerical implementation of the method.* The simplex method involves several computational schemes for calculating, storing and transforming the matrices and vectors. In the basic version (the algorithm with the inverse matrix) the matrix  $A_1^{-1}$  is computed and stored, whereas expressions such as  $A_2^T(A_1^{-1})^T c^1$  (see (9)) are calculated through a multiplication of this matrix by the corresponding vectors. However, in computing the matrix  $A_1^{-1}$ , the matrices  $A_1$  in successive iterations obtain if only one column has changed. Hence these matrices can be inverted recursively, using relations similar to Lemmas 3 and 4 of Section 3.3. The initial approximation for  $A_1^{-1} b$  in the artificial basis method (12) as well as in problem (14) requires no matrix inversion, since in this case  $A_1 = I$ .

No elaborate formulas will be given for the simplex method. The simplex method software is readily available, and is very sophisticated. There will be hardly many readers who would need to solve linear programming problems manually, or to program the simplex method on their own (the literature on linear programming is plentiful).

3. *Singularity of the simplex method.* We assume that all of the vertices are nonsingular. It would be impossible to invert the matrix  $A_1$  without this assumption. The implementation of the simplex method has, however, shown that in most practical problems the singularity occurs rarely, and therefore one can easily ignore it.

### 10.2.3 Other Finite Methods

As was mentioned earlier, there are many computational schemes for the simplex method. We described the one involving the *algorithm with the inverse matrix*. In the revised simplex method, the matrix  $A_1^{-1} A_2$ , rather than the matrix  $A_1^{-1}$ , is stored and updated. Also, there are so-called *product forms* of each method, where the corresponding matrices are not stored in the final form but, rather, computed anew by using elementary transformations corresponding to the preceding iterations.

The appropriate choice of a particular modification depends on the relation between  $m$  and  $n$ , the sparsity of the constraint matrix (in most large-scale problems the matrix  $A$  contains many zeros), the computer memory, the requirements regarding the accuracy of solution, and so on. Still, if one ignores roundoff errors, all implementations of the simplex method proved to give the same sequence of approximations  $x^k$ .

At the same time, there are methods which are based on the notion of the simplex method but are yet different from it. As was noted (see Exercise 2), the simplex method yields solutions to the dual problem as well. Thus, the quantity  $y^k = (A_1^{-1})^T c^1$  can be viewed as an approximation to a solution of the dual problem (which is not admissible for the dual problem, as  $(b, y^k) > (b, y^*)$  until the last iteration, where  $y^k = y^*$  (see Exercise 3)). Hence the following way of solving (1) is also possible. Applying the simplex method to the dual problem, we simultaneously obtain approximations for the primal variables  $x^k$ , which furnish a lower bound for the objective function:  $(c, x^k) \leq (c, x^*)$ . The computations can be such that in solving the dual problem the basic process is the iterative process for the primal variables. This technique is known as the *dual simplex method*. It is convenient to apply this method in the following cases. For instance, if the primal problem has the form (1) of Section 10.1, the dual problem is written in canonical form, making thereby the simplex method applicable without additional transformations. Another case involves solving a sequence of linear programming problems, each of which is obtained from the preceding problem by adding a new inequality constraint (see, e.g., the cutting plane method and other methods for solving nonsmooth problems in Section 5.4). Then the solution of the preceding problem is not an admissible point for the new problem, and hence it cannot be used in the simplex method. However, the solution of the dual problem turns out to be a vertex for the new dual problem, and it is a good approximation for the dual simplex method (see Exercise 4).

There are some other finite methods which are based on the notion of the simplex method. For example, a method for solving simultaneously the primal and the dual problems. This method is convenient because it derives the upper and the lower bounds for the objective function. In this case the iterations can be terminated as soon as an approximation is obtained with the desired accuracy. The fact is, however, that these methods are not used widely, which can be explained, perhaps, by a successful program implementation of the simplex method, making thus any other method of linear programming unnecessary.

### Exercises

3. Show that for the quantity  $y^k = (A_1^{-1})^T c^1$  in the simplex method the relation  $(b, y^k) > (b, y^*)$  holds iff the  $x^k$  is not a solution of problem (1).

4. Let  $x^*$  be a solution of the problem  $\min(c, x)$ ,  $Ax \leq b$ , and let  $y^*$  be a solution of the dual problem. Write the problem dual to the problem  $\min(c, x)$ ,  $Ax \leq b$ ,  $(a^{m+1}, x) \leq b_{m+1}$ , and show that  $\{y^*, 0\} \in \mathbf{R}^{m+1}$  is a vertex for this problem.

#### 10.2.4 Why Does the Simplex Method Work?

All of the foregoing arguments do not allow us to conclude that the simplex method is an efficient method for solving linear programming problems of any large dimension. Moreover, this method might be inoperable for such problems. Indeed, the only fact we know about the convergence of the method is that it is finite. But how many steps does it take for the method to converge? We can guarantee *a priori* that the number of steps should not exceed the number of vertices of the admissible set (denote this number by  $N$ ). It follows from Lemma 1 of Section 10.1 that the bound  $N \leq C_m^n$  (we refer to constraints of the form (13) in Section 10.1) holds. If  $m = 2n$ , then by Stirling's formula,  $C_m^n \sim 2^{2n} (n/e)^n$  is an astronomical quantity even when  $n$  is of the order less than a hundred. This is, of course, only the upper bound for  $N$ , and is significantly exaggerated; however, it is rather gloomy. As simple examples illustrate, the number of vertices of a polyhedron is large indeed (although it is essentially smaller than  $C_m^n$ ). For instance, let  $Q = \{x \in \mathbf{R}^n : 0 \leq x \leq a\}$ ,  $a > 0$ . Then obviously  $N = 2^n$ . For  $n = 100$  (this is a problem of small dimension in linear programming) we have  $N = 2^{100} \approx 10^{30}$ , which is beyond the capability of any currently available computer. There considerations concern only the number of vertices of a polyhedron and not directly the number of steps of the simplex method. However, special examples have been constructed in which a trajectory of the simplex method passes over all the vertices of the constraints polyhedron in [10.18]; cf. also problem 11 in Section 12.3: the number of steps increases exponentially as the dimension increases.

These considerations seem to create a pessimistic picture of what the simplex method can do. George Dantzig, the creator of the simplex method, writes in [10.5] that the simplex method was first rejected as inefficient, and only by luck was it confirmed and accepted. Contrary to all the expectations, a numerical study demonstrated a remarkably high efficiency of this method in solving practical problems. For the majority of problems the number of iterations is of the order  $m\sqrt{2m}$  (viz. the problem in canonical form (1)) and does not grow exponentially as the dimension increases. Only rarely (thousands problems have been solved through the simplex method) was the number of iterations significantly larger than  $m$ . The reason for such amazing "luck" remains obscure.

A probable explanation comes from the investigation of an "average" behavior of the simplex method. Let us define the meaning of this word by studying the average number of vertices of a polyhedron.

I ...

Consider various polyhedra described by inequalities

$$(a^i, x) \leq \alpha_i b_i, \quad i = 1, \dots, m, \quad (15)$$

where  $a^i \in \mathbf{R}^n$ ,  $b_i \in \mathbf{R}^1$  are fixed, and  $\alpha_i$  run thru  $\pm 1$ . The condition for a *general position* is assumed to be satisfied: any  $n$  of vectors  $a^i$  are linearly independent and the number of active constraints at any point does not exceed  $n$ . Then the total number of polyhedral sets defined by (15) for various  $a^i = \pm 1$  is

$$P = 2^m.$$

The total number of vertices for all these polyhedra is

$$V = C_m^n$$

(by Lemma 1 a vertex is determined by  $n$  active constraints). Each vertex belongs to  $2^n$  polyhedral sets. Hence the average number of vertices is

$$a(n, m) = \frac{2^n V}{P} = 2^{n-m} C_m^n. \quad (16)$$

The term “average” can be interpreted also in the following probabilistic sense. Let the vectors  $\{a^i, b_i\} \in \mathbf{R}^{n+1}$  be random with distribution symmetrical under changing sign of  $b_i$  and such that the condition for a general position holds with probability 1. Then the expected number of vertices for a polyhedral set

$$(a^i, x) \leq b_i, \quad i = 1, \dots, m,$$

is determined by (16).

It follows from (16) that if  $n$  is fixed and  $m \rightarrow \infty$ , then  $a(n, m) \rightarrow 0$ . It means that the “average” polyhedron will be empty. Thus the problem of an average number of vertices of a nonempty polyhedron is of interest. Denote the total number of nonempty polyhedra determined by (15) as  $N(n, m)$ . The following recurrence relation holds:

$$N(n, m+1) = N(n, m) + N(n-1, m).$$

Indeed, an additional hyperplane  $(a^{m+1}, x) = b_{m+1}$  generates as many additional polyhedral sets in  $\mathbf{R}^n$  as the number of such sets in the hyperplane (i.e., in  $\mathbf{R}^{n-1}$ ). The solution of this recurrence relation subject to the obvious boundary values

$$N(1, m) = m + 1, \quad N(n, 1) = 2,$$

is given by

$$N(n, m) = \sum_{k=0}^n C_m^n. \quad (17)$$

In other words,  $m$  hyperplanes of a general position divide  $\mathbf{R}^n$  into  $N(n, m)$  polyhedral parts. Hence the average number of vertices of a nonempty polyhedron is

$$\bar{\alpha}(n, m) = \frac{2^n V}{N(n, m)} = \frac{2^n}{\sum_{k=0}^n C_m^k}. \quad (18)$$

If  $n$  is fixed and  $m \rightarrow \infty$ , it then follows from (18) that

$$\bar{\alpha}(n) = \lim_{m \rightarrow \infty} \bar{\alpha}(n, m) = 2^n. \quad (19)$$

Thus, the average number of vertices does not increase when the number of constraints increases but the dimension of the space is fixed. Thus the average performance of the number of vertices may be quite different from that in the worst case.

Investigations of such kind (but far more sophisticated) can be applied to the problem of evaluating the number of steps for the simplex method. It has been established that the average number of steps does not increase exponentially with  $m$  and  $n$ . The strongest results verified the assumptions made above that this number depends linearly on the dimension of the problem. It means that “bad” problems (i.e., which require a great number of steps) are rare.

### Exercises

5. Write the simplex method for the problem  $\min (c, x)$ ,  $0 \leq x_i \leq 1$ ,  $i = 1, \dots, n$ . Show that it always terminates in no more than  $n$  steps.
6. Calculate the average number of edges, faces and bounded polyhedra among  $N(n, m)$ .

## 10.3 ITERATIVE METHODS OF LINEAR PROGRAMMING

### 10.3.1 The Need For Iterative Methods

In Section 10.2 we reviewed the advantages of the simplex method. However, this method has its limitations—for example, it is sensitive to roundoff errors, which for ill-posed problems can result in a significant loss of accuracy (their effect has not been fully investigated yet, either in theory or in practice). Furthermore, in using the simplex methods one needs to store the coefficients of the inverse matrix  $A_1^{-1}$ , or

some equivalent quantities. More elaborate programs are specially devised as to minimize the storage, but it is still enormous for problems of large dimensions. Neither, one can use the *a priori* information about the problem, such as a good initial approximation for the primal and/or dual variables. Finally, most routines of the simplex method are not for use as subroutines, whereas in methods of nonlinear programming and of nosmooth optimization a linear programming problem is to be solved in each iteration. Thus there arises the need for new methods of linear programming.

### 10.3.2 Iterative Finite Methods

A comparison of the iterative methods versus the finite methods is not definitive: the simplex method has a definite iteration structure, and the property that the method is finite is given by particular conditions of the problem. We shall examine next several iterative methods which were considered earlier regarding nonlinear problems. For the linear programming problem, those methods terminate in a finite number of steps.

To begin, we consider two methods which are of theoretical rather than computational interest (auxiliary problems solved at each step are not simpler than the initial one).

1. **The regularization method.** Instead of the initial problem

$$\min (c, x), \quad Ax \leq b, \quad (1)$$

we solve the sequence of regularized problems

$$\min [(c, x) + \varepsilon_k \|x\|^2], \quad Ax \leq b. \quad (2)$$

A solution of this problem exists (if the admissible set is nonempty) and is unique; we denote it by  $x^k$ . It follows from Theorem 12 of Section 10.1 that the method is finite.

**THEOREM 1.** Let the set of solutions  $X^*$  of (1) be nonempty. Then there exists an  $\bar{\varepsilon} > 0$  such that for all  $0 < \varepsilon_k < \bar{\varepsilon}$  one has  $x^k = x^*$ , where  $x^*$  is the normal solution of the initial problem (i.e. the element of  $X^*$  with smallest norm).  $\square$

Thus, in the regularization method for linear programming problems, in contrast to the general case (see Theorem 4 in Section 6.1), there is no need to make the regularization parameter tend to 0, since the solution does not change for all sufficiently small values of  $\varepsilon_k$ .

2. **The Gradient projection method.** We construct, as in the prox-method (see Section 6.1), the auxiliary problem in the

$k$ th iteration of the form

$$\min [(c, x) + (2\gamma)^{-1} \|x - x^k\|^2], \quad Ax \leq b, \quad (3)$$

and take its solution for  $x^{k+1}$ . Obviously, this method can be written as the gradient projection method

$$x^{k+1} = P_Q(x^k - \gamma c), \quad Q = \{x: Ax \leq b\}. \quad (4)$$

**THEOREM 2.** If the solution  $x^*$  of problem (1) exists and is unique, then for any  $\gamma > 0$  method (3) is finite, i.e.,  $x^k = x^*$  for some  $k$ .

Indeed, a sharp minimum is attained in (1) (see Lemma 5 in Section 10.1), and according to the general Theorem 1 of Section 7.2, the gradient projection method is finite for a sharp minimum.  $\square$

Method (3) is, of course, ineffective: instead of the initial problem a sequence of more complicated auxiliary problems is solved. The method can be improved if only the  $\varepsilon$ -active constraints are taken into account at each step, i.e., if the problem

$$\begin{aligned} & \min [(c, x) + (2\gamma)^{-1} \|x - x^k\|^2], \\ & (a^i, x) \leq b_i, \quad i \in I_k, \\ & I_k = \{i: (a^i, x^k) \geq b_i - \varepsilon\} \end{aligned} \quad (5)$$

is solved.

More advantageous is however the transition to the dual methods for solving auxiliary problems. We shall describe it next.

**3. The penalty function method.** Instead of the quadratic problem (2) in the regularization method, we shall solve its dual. By (12) and (13) of Section 10.4, the point  $x^k$ , the solution of (2), can be found from the formula

$$x^k = -(2\varepsilon_k)^{-1}(c + A^T y^k), \quad (6)$$

where  $y^k$  is the solution of the problem

$$\min_{y \geq 0} \left[ (b, y) + \frac{1}{4\varepsilon_k} \|c + A^T y\|^2 \right]. \quad (7)$$

Problem (7) reduces to the problem of minimizing a quadratic function on the nonnegative orthant; the minimum can be found through the conjugate gradient method (Section 7.3). On the other hand, (7) is nothing but the penalty function method for solving the problem dual of (1):

$$\begin{aligned} & \min (b, y), \\ A^T y &= -c, \quad y \geq 0, \end{aligned} \tag{8}$$

where one takes  $1/(4\epsilon_k)$ ,  $\epsilon_k \rightarrow 0$ , as penalty coefficient. Thus the regularization method is equivalent to the penalty function method for the dual problem.

Conversely, one can use the penalty function method for the primal problem

$$\begin{aligned} & \min (c, x), \\ Ax &= b, \quad x \geq 0, \end{aligned} \tag{9}$$

by solving a sequence of quadratic minimization problems on  $\mathbf{R}_+^n$ :

$$\min_{x \geq 0} [(c, x) + \frac{1}{2} K_k \|Ax - b\|^2], \quad K_k \rightarrow \infty. \tag{10}$$

We denote the solution of this problem by  $x^k$  and use the relationship with the regularization method, as well as Theorem 1, to obtain the following theorem.

**THEOREM 3.** If a solution of problem (9) exists, then for sufficiently large  $K_k$  method (10) yields an exact solution for the dual problem:

$$y^* = K_k(Ax^k - b). \quad \square \tag{11}$$

Thus, the penalty function method for linear programming problems is finite for the dual problem (although  $x^k$  is not necessarily the solution of the primal problem).

4. The augmented Lagrangian method. We write the problem dual to (3) and obtain the method

$$\begin{aligned} y^k &= \underset{y \geq 0}{\operatorname{argmin}} \left[ (b, y) - (A^T y, x^k) + \frac{1}{2} \gamma \|A^T y + c\|^2 \right], \\ x^{k+1} &= x^k - \gamma(A^T y^k + c). \end{aligned} \tag{12}$$

This is exactly the augmented Lagrangian method for the dual problem (8) with  $K = \gamma$ . Conversely, if the initial problem has canonical form (9), then the augmented Lagrangian method

$$\begin{aligned} x^k &= \underset{x \geq 0}{\operatorname{argmin}} M(x, y^k, K), \\ M(x, y, K) &= (c, x) + (y, Ax - b) + \frac{1}{2} K \|Ax - b\|^2, \end{aligned} \tag{13}$$

$$y^{k+1} = y^k + K(Ax^k - b)$$

is equivalent to the gradient projection method for the problem dual to (9) with  $\gamma = K$ . Noting Theorem 2 and Exercise 1, we have the following assertion.

**THEOREM 4.** If a solution of problem (9) exists, then for any  $K > 0$  method (13) yields a solution to both the primal and the dual problems in a finite number of steps.  $\square$

The augmented Lagrangian method (13) has many attractive features. Since it does not require  $K$  to be too large, the auxiliary problems in (13) are not necessarily ill-posed. On the other hand, as  $K$  increases the number of iterations decreases and, in principle, can be reduced to 1 (see Exercise 4). A solution to problem (13) can be found in a finite number of steps by the conjugate gradient method. Note that there is no need to solve the auxiliary problems (13) exactly; the results concerning the convergence can be obtained for the case, too, where the minimization in (13) is approximate, with adjustable accuracy. Finally, as in most of the iterative methods, in (13) one has to deal only with the initial constraint matrix  $A$ , which permits using its sparsity. These favorable features contributed to its wide applicability in numerical analysis. In many cases the augmented Lagrangian method is as good as the simplex method.

### Exercises

1. Prove Theorem 2 without assuming that the solution is unique.
2. Suppose the primal problem has the form (1) and the penalty function method is used:

$$x^k = \operatorname{argmin} [(c, x) + (K_k/2) \| (Ax - b)_+ \|^2]$$

V ]

Prove that for sufficiently large  $K_k$  the vector  $K_k(Ax^k - b)_+$  is a solution of the dual problem (8).

3. Prove that if the admissible set in (1) is bounded, then we can find a  $\bar{\gamma} > 0$  such that for all  $x^0 \in Q$  and all  $\gamma \geq \bar{\gamma}$ , method (3) yields a solution in a single step, i.e.,  $x^1 = x^*$ .
4. Using the preceding result, show that for sufficiently large  $K$  method (13) yields a solution in a single step.

#### 10.3.3 Reduction to Nonsmooth Minimization

Although advantageous, the above methods cannot be regarded to be a universal tool for solving arbitrary linear programming problems. Above all, they require that a rather complex auxiliary problem be solved at each step, and for large-scale problems the computation becomes too laborious or even unfeasible. Hence it is worth considering other iterative methods, in which

the computations in a particular iteration are very simple (perhaps, by slowing down the convergence rate). These methods make it possible to find (although not exactly) an approximate solution of a large-scale problem.

The first approach of this kind involves simple methods of nonsmooth optimization (see Chapter 5). The primal linear programming problem can be reduced to unconstrained minimization of a piecewise linear function in various ways. For example, let

$$f(x) = (c, x) + K \max_{1 \leq i \leq m} [(a^i, x) - b_i]_+ . \quad (14)$$

Then for sufficiently large  $K \geq \bar{K}$  the minimum points of  $f(x)$  on  $\mathbf{R}^n$  coincide with the set of solutions  $X^*$  of the problem

$$\begin{aligned} & \min (c, x), \\ & (a^i, x) \leq b_i, \quad i = 1, \dots, m . \end{aligned} \quad (15)$$

Note that a similar method of nonsmooth penalties has already been discussed in regard to convex programming problems (Lemma 2 of Section 9.3). One of the subgradients of (14) is computed very easily:

$$\partial f(x) = c + K a^j, \quad j = \operatorname{argmax}_{1 \leq i \leq m} [(a^i, x) - b_i]_+, \quad (16)$$

i.e., it suffices to find the “most violated” constraint at  $x$ . The drawback of this approach is that  $\bar{K}$  is usually unknown.

A very simple way of transforming the problem is to use the dual function. Suppose in the primal problem there are two-sided constraints:

$$\begin{aligned} & \min (c, x), \quad x \in \mathbf{R}^n, \\ & Ax = b, \quad b \in \mathbf{R}^m, \quad a \leq x \leq d . \end{aligned} \quad (17)$$

Introduce the dual function (see (47) of Section 9.1):

$$\theta(y) = - \min_{a \leq x \leq d} [(c, x) + (y, Ax - b)] . \quad (18)$$

For every  $y$  a solution  $x(y)$  of the minimization problem in  $x$  is very simple to find:

$$x(y)_i = \begin{cases} a_i & \text{if } (c + A^T y)_i > 0, \\ d_i & \text{if } (c + A^T y)_i < 0 . \end{cases} \quad (19)$$

The function  $\theta(y)$  is convex, piecewise linear, everywhere defined, and by the duality theorem the primal problem is equivalent to unconstrained

minimization of  $\theta(y)$ . A subgradient of  $\theta(y)$  has the form

$$\partial\theta(y^k) = -(Ax^k - b), \quad x^k = x(y^k). \quad (20)$$

This approach is most appropriate when the dimension of the dual variables is small:  $m \ll n$ .

Finally, we write the extremum conditions for (15) (see Theorem 3 of Section 10.1) in the form of equalities and inequalities:

$$Ax \leq b, \quad (c, x) = (b, y), \quad y \geq 0, \quad A^T y - c = 0, \quad (21)$$

and reduce the linear programming problem to that of minimizing the residuals in (21):

$$\phi(x, y) = \|(Ax - b)_+\| + |(c, x) - (b, y)| + \|y_+\| + \|A^T y - c\|. \quad (22)$$

This method is not efficient since the problem has dimension  $n+m$ . The advantage of this method is that the minimal value of  $\phi$  is known:  $\phi^* = 0$ .

Upon reduction of this problem to that of unconstrained minimization (for the sake of definiteness, say, by (14)) we apply the subgradient method:

$$x^{k+1} = x^k - \gamma_k \partial f(x^k). \quad (23)$$

Under the usual conditions on  $\gamma_k$ , the results obtained in Section 5.3 ensure convergence. If the solution  $x^*$  is unique, then  $f(x)$  has a sharp minimum and therefore the estimates of the convergence rate obtained therein hold also for a sharp minimum.

Instead of the subgradient method (23) one can use more powerful methods for increasing the convergence described in Section 5.4, e.g., the ellipsoid method (12) or the space dilation method (14) therein. Such methods have limited capabilities, since they require storage of a matrix  $H_k$  of large dimension, with the convergence rate being not too great (see Theorem 4 of Section 5.4). In this connection, Khachiyan [10.15] derived an interesting theoretical result, viz. it is possible to find an exact solution in a finite number of arithmetic operations, which depends polynomially on the dimension of the problem (we refer the reader to [10.15] for the precise statement; also see Exercises 5 and 6). The estimate in Theorem 4 of Section 5.4 plays a key role in proving this result: to attain a specified accuracy with respect to the function, one must have roughly  $O(n^2)$  iterations of method (12) of Section 5.4. On the other hand, for the simplex method and other finite methods there are only exponential estimates of the number of operations (as was noted in Section 10.2, there are cases where the simplex method requires the number of operations approximately of the order  $\exp(O(n))$ ). One should not infer, however, that the iterative method based

✓ )

on a reduction of a linear programming problem to unconstrained minimization together with an application of the ellipsoid method is superior to the simplex method for large-scale problems: (a) for finite  $n$ , it is impossible to determine which is larger:  $O(n^0)$  or  $\exp(O(n))$ , since it depends on the unknown constants as well as the value of the variable  $n$  in the formulas, and (b) (which is the key point) the estimates of the simplex method are computed for the “worst” problems. However, most of the real practical problems can be solved, as was noted earlier, in approximately  $m/2m$  iterations. At the same time, estimation of the convergence rate in the case of the ellipsoid method is feasible for any problem (see Section 5.4). Lastly, exactly how effective the method depends on many factors: not only on the number of operations but also on the stability toward roundoff errors, needed computer memory, and others.

### Exercises

5. Suppose the system of inequalities is given:  $\sum_{j=1}^n a_{ij}x_j \leq b_i$ ,  $i = 1, \dots, m$ , where  $a_{ij}$ ,  $b_i$  are integers,  $|a_{ij}| \leq h$ ,  $|b_i| \leq h$ . Also let  $L = n \log_2 h\sqrt{n} + \log_2(n+1)$ . Prove that if the system has a solution, then we can find a solution  $x^*$  such that  $\|x^*\| \leq 2^L$ ; otherwise we have  $\min f(x) \geq 2^{-L}$ , where

$$f(x) = \max_{1 \leq i \leq m} \left( \sum_{j=1}^n a_{ij}x_j - b_i \right).$$

6. Using the preceding result, show that method (12) of Section 5.4 for minimizing  $f(x)$ , indicates whether the system of inequalities has a solution or not in  $O(n^2 L)$  iterations.

#### 10.3.4 The Lagrange Functions

In Section 9.3 we described the dual methods of convex programming, which are used to find a saddle point for the Lagrange function. The simplest method, the Arrow-Hurwicz-Uzawa method, does not converge (see, e.g., (14) in Section 9.3). However, if the Lagrangian is replaced by the augmented Lagrangian, the new method (see (20) of Section 9.3) converges. Thus, for problem (1) consider the method

$$\begin{aligned} x^{k+1} &= x^k - \gamma M'_x(x^k, y^k, K) = x^k - \gamma(c + A^T[y^k + K(Ax^k - b)]_+) , \\ y^{k+1} &= y^k + \gamma M'_y(x^k, y^k, K) = y^k + \gamma \max \{-y^k/K, Ax^k - b\} \end{aligned} \quad (24)$$

and for problem (9) consider the method

$$\begin{aligned} x^{k+1} &= [x^k - \gamma M_x'(x^k, y^k, K)]_+ = [x^k - \gamma(c + A^T y^k + KA^T(Ax^k - b))]_+, \\ y^{k+1} &= y^k + \gamma M_y'(x^k, y^k, K) = y^k + \gamma(Ax^k - b). \end{aligned} \quad (25)$$

**THEOREM 5.** If a solution of problems (1) or (9) exists, then for every  $K > 0$  we can find a  $\bar{\gamma} > 0$  such that for  $0 < \gamma < \bar{\gamma}$  methods (24) and (25) converge to a solution of the corresponding primal and dual problems with the rate of geometric progression.  $\square$

It is worth noting that the progression ratio cannot be made small by the choice of  $K$ , so that the convergence rate cannot be very large.

In the Arrow-Hurwicz-Uzawa method, the iteration consists of a gradient step of minimization in  $x$  and of maximization in  $y$  of the Lagrange function. Instead, one can alternate finding the minimum in  $x$  and the minimum in  $y$  of the function, and shift in the direction thus obtained. Say, we have the problem

$$\begin{aligned} \min (c, x), \quad x \in \mathbf{R}^n, \\ Ax \leq b, \quad b \in \mathbf{R}^m, \\ x \geq 0, \quad c > 0, \quad b > 0. \end{aligned} \quad (26)$$

Furthermore, let  $x^0$  be some admissible point,  $(c, x^0) = \alpha$ . Then the solution of problem (26) does not change if we add the constraint  $(c, x) \leq \alpha$ ; likewise, by the duality theorem every solution of the dual problem must satisfy the condition  $(b, y) \leq \alpha$ . Hence (26) can be reduced to the following saddle-point problem:

$$\begin{aligned} \min_{x \in Q} \max_{y \in S} L(x, y) &\equiv \max_{y \in S} \min_{x \in Q} L(x, y), \\ L(x, y) &= (c, x) + (y, Ax - b), \\ Q &= \{x \geq 0, (c, x) \leq \alpha\}, \quad S = \{y \geq 0, (b, y) \leq \alpha\}. \end{aligned} \quad (27)$$

We shall be seeking recursively the saddle point by minimizing  $L(x, y^k)$  in  $x \in Q$  and maximizing  $L(x^k, y)$  in  $y \in S$ :

$$\begin{aligned} x^{k+1} &= x^k + \gamma_k(\bar{x}^k - x^k), \quad \bar{x}^k = \operatorname{Argmin}_{x \in Q} L(x, y^k), \\ y^{k+1} &= y^k + \gamma_k(\bar{y}^k - y^k), \quad \bar{y}^k = \operatorname{Argmax}_{y \in S} L(x^k, y). \end{aligned} \quad (28)$$

If  $b > 0$ ,  $c > 0$ ,  $\alpha > 0$ , the points  $\bar{x}^k$ ,  $\bar{y}^k$  are found in a simple way:

$$\bar{x}^k = \begin{cases} 0, & \text{if } c + A^T y^k \geq 0, \\ \frac{\alpha}{(c + A^T y^k)_j} e_j, & j = \underset{1 \leq i \leq m}{\operatorname{argmin}} \frac{(c + A^T y^k)_i}{\alpha} \text{ otherwise;} \end{cases} \quad (29)$$

$$\bar{y}^k = \begin{cases} 0, & \text{if } Ax^k - b \geq 0, \\ \frac{\alpha}{(Ax^k - b)_j} e_j, & j = \underset{1 \leq i \leq m}{\operatorname{argmax}} \frac{(Ax^k - b)_i}{\alpha} \text{ otherwise.} \end{cases}$$

**THEOREM 6.** If in problem (26)  $c > 0$ ,  $b > 0$ ,  $\alpha > 0$ , then in method (28), (29) for

$$\gamma_k \rightarrow 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty, \quad (30)$$

all limit points make a solution of the *primal* and the *dual* problems.  $\square$

Method (28), (29) was originally developed in game theory as the so-called *fictitious play* method or the Brown-Robinson method. It is an extremely simple method. Unfortunately, in theory as well as in numerical computations this method proved to be slowly convergent. Many heuristic attempts to increase the rate of convergence have not been successful.

### 10.3.5 Summary

Until recently, the simplex method was viewed as a universal and efficient tool for solving linear programming problems, and there hardly been a need for any other methods, especially, iterative methods. Today, there are other opinions, too. Computational schemes of iterative methods to match the simplex method are in order.

## 10.4 QUADRATIC PROGRAMMING

We say that the problem of minimizing a quadratic form under the linear constraints

$$\begin{aligned} \min & [(Cx, x)/2 - (d, x)], \quad x \in Q, \\ Q &= \{x: Ax \leq b\} \end{aligned} \quad (1)$$

is a *quadratic programming problem*. Here  $x \in \mathbf{R}^n$ ,  $C$  is a symmetric  $n \times n$  matrix,  $b \in \mathbf{R}^m$ ,  $A$  is an  $m \times n$  matrix. We know that the quadratic form

$(Cx/x)/2 - (d, x)$  is convex iff  $C \geq 0$ . Without this condition, it is, in general, a multimodal problem. In what follows we assume that

$$C \geq 0. \quad (2)$$

Note that the linear constraints in (1) can be transformed by the same methods as those in linear programming problems, so we be using the most convenient notation.

#### 10.4.1 Extremum Conditions

To specify the general extremum conditions for convex problems (Theorem 2 of Section 9.1), we have the following theorem.

**THEOREM 1.** A necessary and sufficient extremum condition for problem (1), (2) at  $x^*, Ax^* \leq b$ , is the existence of a  $y^* \in \mathbf{R}^m$  such that

$$Cx^* - d + A^T y^* = 0, \quad y^* \geq 0, \quad y^*(Ax^* - b) = 0. \quad \square \quad (3)$$

In order to write the dual problem for (1), we form the Lagrangian

$$L(x, y) = (Cx, x)/2 - (d, x) + (y, Ax - b) \quad (4)$$

and the dual function

$$\psi(y) = \inf_x L(x, y). \quad (5)$$

Let us define the form of  $D(\psi)$ , i.e., the domain where  $\psi(y) \neq -\infty$ . If a quadratic function is bounded from below on  $\mathbf{R}^n$ , it attains a minimum (see Exercise 2 in Section 1.3). Hence, for any  $y \in D(\psi)$  we can find an  $x = x(y)$  such that  $L'_x(x, y) = 0$ , i.e.,

$$Cx - d + A^T y = 0.$$

Here

$$\begin{aligned} \psi(y) &= L(x(y), y) = (Cx, x)/2 + (A^T y - d, x) - (b, y) \\ &= -(Cx, x)/2 - (b, y). \end{aligned}$$

Since the dual problem has the form (49) of Section 8.1:

$$\min \theta(y),$$

$$\theta(y) = -\psi(y), \quad y \geq 0, \quad y \in D(\theta),$$

the solving of it is equivalent to that of the problem

$$\begin{aligned} \min & [(Cx, x)/2 + (b, y)] , \\ Cx + A^T y &= d . \end{aligned} \quad (6)$$

We can exclude  $x$  in two important cases of (6): for  $C = 0$  (a linear programming problem) and for  $C > 0$ :

$$x = C^{-1}(d - A^T y) , \quad (7)$$

where  $D(\psi) = \mathbf{R}^m$ . We obtain the dual problem

$$\begin{aligned} \min & [\frac{1}{2}(C^{-1}(d - A^T y), (d - A^T y)) + (b, y)] . \\ y \geq 0 & \end{aligned} \quad (8)$$

Thus for a pair of dual problems we have the following theorem.

**THEOREM 2** (duality theorem). For  $C > 0$  problem (8) is dual to problem (1): their solutions  $x^*$ ,  $y^*$  exist, or do not exist, simultaneously; they are related through

$$x^* = C^{-1}(d - A^T y^*) , \quad (9)$$

and the inequalities

$$(Cx, x)/2 - (d, x) \geq (C^{-1}(d - A^T y), d - A^T y)/2 + (b, y) \quad (10)$$

are satisfied for all  $Ax \leq b$ ,  $y \geq 0$ ; equality is possible only for  $x = x^*$ ,  $y = y^*$ .  $\square$

Thus for  $C > 0$  the quadratic programming problem reduces to that of minimizing a quadratic function on  $\mathbf{R}_+^m$ . We have used this reduction more than once in studying iterative methods at each step of which a quadratic programming problem was being solved. The dual problem is especially simple when  $C = \gamma^{-1}I$ , i.e., if the primal problem has the form

$$\min [(d, x) + (2\gamma)^{-1} \|x - a\|^2] , \quad Ax \leq b . \quad (11)$$

Construction of its dual does not require matrix inversion:

$$\min (\gamma/2) \|d + A^T y\|^2 - (y, Ax - b) , \quad y \geq 0 , \quad (12)$$

and from Theorem 2 we find that the solutions  $x^*$ ,  $y^*$  of problem (11), (12) are related through

$$x^* = a - \gamma(A^T y^* + d) . \quad (13)$$

## Exercises

1. Verify that any vertex of the polyhedron  $Q = \{x \in \mathbf{R}^n : |x_i| \leq 1, i = 1, \dots, n\}$  is a local minimum point of the function  $f(x) = -\|x\|^2$ .

2. Verify that for the problems

- (a)  $\min [(Cx, x)/2 - (d, x)], x \geq 0;$
- (b)  $\min [(Cx, x)/2 - (d, x)], Ax = b;$
- (c)  $\min \|x - a\|^2, Ax = b;$
- (d)  $\min [(Cx, x)/2 - (d, x)], Ax = b, x \geq 0,$

the extremum conditions have the following form:

- (a)  $Cx^* - d \geq 0, (Cx^* - d, x^*) = 0;$
- (b)  $Cx^* - d + A^T y^* = 0$  (cf. Section 7.1);
- (c)  $x^* - a + A^T y^* = 0;$
- (d)  $Cx^* - d + A^T y^* \geq 0, (Cx^* - d + A^T y^*, x^*) = 0.$

### 10.4.2 Existence, Uniqueness and Stability of a Solution

The case  $C > 0$  is very simple to investigate.

**THEOREM 3.** Let  $Q$  be nonempty,  $C > 0$ . Then a solution  $x^*$  of problem (1) exists, is unique and

$$f(x) - f(x^*) \geq \alpha \|x - x^*\|^2, \quad \alpha > 0 \quad (14)$$

is satisfied for all  $x \in Q$ .

Indeed, in this case the objective function is strongly convex and one can apply Exercise 6 of Section 7.1.  $\square$

The case  $C = 0$  has been analyzed earlier. The quadratic programming problem becomes a linear programming problem.

For  $C \geq 0$  the results are very similar.

**THEOREM 4.** Let  $C \geq 0$ , let  $Q$  be nonempty, and let  $(Cx, x)/2 - (d, x)$  be bounded from below on  $Q$ . Then a solution of problem (1) exists.

The proof of this theorem follows the lines of the proof of Theorem 9 in Section 10.1.  $\square$

**COROLLARY.** The problem

$$\begin{aligned} & \min \|Cx - d\|^2, \\ & x \in Q = \{x : Ax \leq b\} \end{aligned} \quad (15)$$

has a solution if  $Q$  is nonempty.  $\square$

The basic results on solution stability follow from general assertions concerning the convex programming problem (see Section 9.1). Here is one specific result on quadratic programming problem.

**THEOREM 5.** Let  $C \geq 0$  and let the set of solutions  $X^*$  of problem (1) be nonempty,  $f^* = f(x^*)$ ,  $x^* \in X^*$ . Then we can find  $\alpha > 0$ ,  $\beta > 0$  such that for any  $x \in Q$ ,

$$\text{either } f(x) - f^* \geq \alpha p^2(x, X^*) \quad \text{or } f(x) - f^* \geq \beta p(x, X^*) . \quad \square \quad (16)$$

### 10.4.3 Finite Methods

It is theoretically possible to solve a quadratic programming problem in a finite number of operations because the minimization of a quadratic function on a linear manifold reduces to solving a system of linear equations and, also, because the number of faces of a polyhedron is finite. Hence, a sequential solution of problems on the faces of the constraint polyhedron can yield a solution of the problem.

One of the possible ways of implementing this technique consists in the following. Assume that we are solving problem (1) for  $C > 0$  and, in addition, at the  $k$ th step we find an admissible point  $x^k$  and determine the set of constraints  $I_k$ , where all  $i \in I_k$  are active. Let us find the solution of the problem

$$\begin{aligned} \min & [(Cx, x)/2 - (d, x)], \\ (a^i, x) &= b_i, \quad i \in I_k . \end{aligned} \quad (17)$$

This problem has a solution by Theorem 3 and the condition  $(a^i, x^k) = b_i$ ,  $i \in I_k$ ; it can be found from the system of linear equations (see Exercise 2(b)):

$$Cx - d + \sum_{i \in I_k} y_i a^i = 0, \quad (a^i, x) = b_i, \quad i \in I_k . \quad (18)$$

If the solutions of this problem,  $\bar{x}^k$ ,  $y_i^k$  are such that  $\bar{x}^k$  satisfies all the constraints  $(a^i, x) \leq b_i$ ,  $i = 1, \dots, m$ , while  $y_i^k \geq 0$ ,  $i \in I_k$ , then  $\bar{x}^k$  is a solution (see Theorem 1). If  $\bar{x}^k$  does not satisfy the constraints, we take

$$x^{k+1} = x^k + \lambda_k (\bar{x}^k - x^k) , \quad (19)$$

where  $0 < \lambda_k < 1$  is chosen from the condition  $\lambda_k = \max \{ \lambda : x^k + \lambda(\bar{x}^k - x^k) \in Q \}$ , we include the new active constraints in  $I_{k+1}$  and repeat the calculation. Finally, if  $\bar{x}^k$  is admissible but some of the  $y_i^k$  are negative, then we take  $x^{k+1} = \bar{x}^k$ ,  $I_{k+1} = \{i \in I_k, y_i^k > 0\}$ . Suppose this procedure starts from the point  $x^0 \in Q$ ,  $I_0 = \{i : (a^i, x^0) = b_i\}$ . Then a solution is found in a finite number of steps.

It is possible to employ the elimination of variables in the same way as in the simplex method. It is however advantageous to state the problem in the canonical form

$$\min [(Cx, x)/2 - (d, x)], \quad Ax = b, \quad x \geq 0. \quad (20)$$

It suffices to require  $C \geq 0$  (rather than  $C > 0$ ). Lastly, introducing sequentially artificial variables makes a computational scheme perfectly analogous to the simplex method schemes. These questions are of limited conceptual interest and we omit their discussion.

#### 10.4.4 Iterative Methods

The dual problem (8) reduces to minimization of a quadratic function on  $\mathbf{R}_+^m$ , which can be achieved by the conjugate gradient method (see Section 7.3). Thus we arrive at the iterative method which leads to a solution in a finite number of iterations. Nevertheless, to compute the gradient, one needs either to construct  $C^{-1}$  or solve at each step a system of linear equations with matrix  $C$ . This is worth doing only when  $C$  is simple (say, the unit matrix).

The finite iterative methods of linear programming in Section 10.3 are no longer finite for quadratic problems. General results on smooth convex problems (see Section 9.3) yield convergence and convergence rate bounds for these methods.

On the whole, for quadratic programming problems general iterative methods of convex programming do not possess any superior efficiency. On the other hand, no detailed iterative algorithms have been developed for such problems. In particular, it still remains unclear how to optimize the use of the notions of the conjugate gradient method for general quadratic programming problems. To conclude, the question of solving large-scale quadratic problems has not been resolved completely.



## **PART III**

## **APPLICATIONS**

## CHAPTER 11

### OPTIMIZATION PROBLEMS: EXAMPLES

Our discussion of minimization problems was somewhat abstract: each chapter had a typical opening like “let us consider the problem of minimizing the function  $f(x)$  on  $\mathbf{R}^n$ ”, assuming that  $f(x)$  is convex and differentiable...” The question is, however, Where do such problems arise? How often do they arise? What is the specific form of the objective function? In what follows we shall illustrate our answers with examples. First we formulate the problem and then examine how to apply the methods described earlier to analyze and solve the problem. We are particularly interested to find out how to combine general methods with specific methods of solving the problem. For many concrete problems we shall suggest special methods, which are more effective than standard ones. Nevertheless, the reader should see for herself/himself that solving any specific problem is not a routine procedure and involves a lot of ingenuity and skill.

We also wish to note that specialists in a particular area of science and technology in which optimization problems arise, developed their own vocabulary and notation system, frequently quite different from the standard terminology used in the theory of extremal problems. Thus, in the literature on Identification it is common to denote unknown parameters by  $a, b, c, \dots$ , and the measurements by  $x, y, \dots$ , so that the notation for a typical problem is  $J(c), J(c) = \sum_{i=1}^m F(x_i - c)$ . We shall be using the conventional language for optimization problems, rather than “local” jargon.

#### 11.1 IDENTIFICATION PROBLEMS

Possibly the need for optimization arises most often in Identification problems, i.e., in constructing a model from the available data. Identification problems exist in many forms, depending on the model—statistical, dynamic, linear, nonlinear, etc.—as well as on assumptions concerning the data—direct, indirect, statistical, deterministic.

### 11.1.1 Statistical Problems of Parameter Estimation

Suppose one has a sequence of measurements  $z^1, \dots, z^m, z^i \in \mathbf{R}^s$ , which is interpreted as an independent sample with distribution density  $p(z, x^*)$ . The form of the density  $p(z, x)$  is known, and the unknown value of the parameter  $x^* \in \mathbf{R}^n$  has to be determined.

For instance, suppose we are given  $m$  measurements of a scalar variable  $x^*$  containing a random error:

$$z^i = x^* + \xi_i, \quad i = 1, \dots, m, \quad (1)$$

where  $\xi_i$  are independent random variables which are normally distributed with mean 0 and variance 1. Then

$$p(z, x) = (2\pi)^{-1/2} \exp(-(z-x)^2/2), \quad (2)$$

and the problem is to find the estimate for  $x^*$  according to  $z^1, \dots, z^m$  and the density (2). A similar problem was treated in Section 4.5.

The most widely used estimation method in this case would be the maximum likelihood method, in which one chooses the estimate for  $x^*$  to maximize the probability of the realization  $z^1, \dots, z^m$ . Since the latter is proportional to  $p(z^1, x) p(z^2, x) \dots p(z^m, x)$  due to the independence of the measurements, the method reduces to maximization of  $p(z^1, x) \dots p(z^m, x)$  in  $x$ . If we take the logarithm of this function (which has no effect on the solution) and reverse the sign, we arrive at the unconstrained minimization problem:

$$\min f_m(x), \quad f_m(x) = -\sum_{i=1}^m \log p(z^i, x). \quad (3)$$

We say that the vector

$$x_m^* = \underset{x \in \mathbf{R}^n}{\operatorname{argmin}} f_m(x) \quad (4)$$

is the maximum likelihood estimate. In particular, for problem (1), (2)

$$f_m(x) = \frac{1}{2} \sum_{i=1}^m (z^i - x)^2 + \frac{1}{2} \log 2\pi$$

and the minimum point of  $f_m(x)$  is found explicitly

$$x_m^* = \frac{1}{m} \sum_{i=1}^m z^i, \quad (5)$$

i.e., the maximum likelihood estimate in this case is the sample mean.

The maximum likelihood method possesses the important property of optimality, which explains why it is widely used. Under certain regularity assumptions on the  $p(z, x)$  the distribution of the quantity  $\sqrt{m}(x_m^* - x^*)$  tends to normal distribution with mean 0 and covariance matrix  $S$ :

$$S = J^{-1}, \quad J = \int \frac{\nabla_x p(z, x^*) \nabla_x^T p(z, x^*)}{p(z, x^*)} dz. \quad (6)$$

$\nearrow$  for  $m \rightarrow \infty$

Here  $J$  is the Fisher information matrix (cf. (23) of Section 4.5). By the Cramer-Rao inequality (see (24) of Section 4.5),  $J^{-1}$  is the lower bound for the covariance matrices of the quantities  $\sqrt{m}(\hat{x}_m - x^*)$ , where  $\hat{x}_m$  is an arbitrary unbiased estimate. Thus the maximum likelihood estimate has the asymptotically best accuracy. In statistical terms, the estimate  $x_m^*$  is asymptotically normal, consistent and asymptotically efficient.

Thus, in order to construct the maximum likelihood estimate, it is necessary to find the unconstrained minimum of  $f_m(x)$  of the form (3). Of course, the  $x_m^*$  can be found explicitly in rare cases, like problem (1), (2). To minimize the  $f_m(x)$ , one usually needs to use numerical methods. A method of minimization depends on several factors: (1) the dimension of the problem is usually small, i.e., the number of parameters sought is rarely larger than 10; (2) optimality of the maximum likelihood estimate is proved under regularity assumptions which include smoothness of  $p(z, x)$  in  $x$ . Hence for problems in which the maximum likelihood method is proven,  $f_m(x)$  is smooth; (3)  $f_m(x)$  is a sum of identical functions (which differ only by the values of  $z^i$ ). Therefore it is not hard to write its second derivatives, and hence the calculation of  $\nabla f_m(x)$  and  $\nabla^2 f_m(x)$  is not difficult; and (4) for many problems (although not for all problems) the function  $f_m(x)$  is convex and thus there is no multimodality. But empirically,  $f_m(x)$  is frequently ill-conditioned.

This discussion as well as the analysis of the methods in Chapters 1 and 3 enables us to conclude that in order to minimize  $f_m(x)$  which has the form (3), it is appropriate to employ Newton's method in combination with the procedures which make this method converge globally (Section 3.1).

### Exercises

- Verify that the maximum likelihood method for estimating the mean and the variance of normal distribution reduces to the minimization of the function

$$f_m(x) = m \log x_2 + \sum_{i=1}^m \frac{(z^i - x_1)^2}{2x_2^2}$$

in  $\mathbf{R}^2$ . Investigate this function with respect to convexity, existence and uniqueness of the minimum, as well as domain of definition. Write its derivatives.

2. Write the maximum likelihood method for the problem of finding the minimum of a quadratic function on the basis of measurements of its gradients (see Subsection 4.5.2).

### 11.1.2 Regression Problems

In the preceding problem all the measurements have the same distribution. A more general situation is the one where the measured variable changes its values depending on some, *a priori* known, input variables. Among these problems, the regression problem is the simplest. Let

$$y_i = \phi(u^i, x^*) + \xi_i, \quad i = 1, \dots, m, \quad (7)$$

where the  $y_i$  are the measurements,  $\phi: \mathbf{R}^s \times \mathbf{R}^n \rightarrow \mathbf{R}^1$  is the known function,  $u^i \in \mathbf{R}^s$  are the input variables,  $x^* \in \mathbf{R}^n$  are the parameters, and  $\xi_i \in \mathbf{R}^1$  are the random errors. For known  $y_i, u^i, i = 1, \dots, m$ , it is required to estimate  $x^*$ . Such problems are frequent in econometrics (renewal of parameters of economic models from statistical data for a period of previous years), in modelling chemical reactors (calculation of parameters of kinetic equations from experimental data), in crystallography (estimation of crystal lattice parameters from X-ray crystallography data), among other applications.

Assume that  $\xi_i$  are random, independent and identically distributed, with known density  $p(z)$ . Then the maximum likelihood estimate has the form

$$x_m^* = \underset{x \in \mathbf{R}^n}{\operatorname{argmin}} f_m(x), \quad f_m(x) = -\sum_{i=1}^m \log p(y_i - \phi(u^i, x)). \quad (8)$$

One can show that under certain assumptions estimate (8) remains asymptotically normal, consistent and asymptotically efficient.

To solve regression problems, other methods are also used, in particular, the least squares method, in which

$$x_m^* = \underset{x \in \mathbf{R}^n}{\operatorname{argmin}} f_m(x), \quad f_m(x) = \sum_{i=1}^m (y_i - \phi(u^i, x))^2. \quad (9)$$

This method (a) does not require knowledge of the noise distribution law and (b) it is simple. For normally distributed noise, the maximum likelihood method coincides with the least squares method. For other distribution laws, the least squares method is not, generally, efficient or asymptotically efficient. However, it has a major advantage. Suppose the model is linear:

$$y_i = (u^i, x^*) + \xi_i, \quad i = 1, \dots, m, \quad x^* \in \mathbf{R}^n, \quad u^i \in \mathbf{R}^n. \quad (10)$$

Then the least squares estimate depends linearly on the measurements:

$$x_m^* = U^+ y , \quad (11)$$

where  $y = \{y_1, \dots, y_m\} \in \mathbf{R}^m$ ,  $U$  is an  $m \times n$  rows  $u^i$ , and  $U^+$  is its pseudo-inverse (Section 6.1). The Gauss-Markov theorem in Statistics asserts that (11) has the smallest covariance matrix among all linear unbiased estimates independently of the noise distribution law.

Thus, the most widely used methods for estimating regression problems have led us, again, to the unconstrained minimization problems (8) or (9). In this case, the functions  $f_m(x)$  are usually smooth; however, problems (8) and (9) are generally nonconvex (for functions  $\phi(u, x)$  which are nonlinear in  $x$ ) and may have a sufficiently large dimension (less than a hundred). The functions in (8) and (9) are composite:

$$f(x) = F(z(x)) , \quad (12)$$

where  $z(x): \mathbf{R}^n \rightarrow \mathbf{R}^m$ ,  $F(z): \mathbf{R}^m \rightarrow \mathbf{R}^1$ . Indeed,

$$z(x)_i = y_i - \phi(u^i, x) \quad \text{and} \quad F(z) = -\sum_{i=1}^m \log p(z_i) ,$$

yields (8), while for  $F(z) = \sum_{i=1}^m z_i^2$  we obtain (9). To minimize functions like (12), it is convenient to apply first-order methods (e.g., the conjugate gradient method, or the quasi-Newton method, see Chapter 3), since the gradient of  $f(x)$  is easy to write:  $\nabla f(x) = z'(x)^T \nabla F(z(x))$ . However, the calculation of  $\nabla^2 f(x)$  is cumbersome since one needs to know the second derivatives of  $z(x)$ , i.e.,  $m$  matrices of dimension  $n \times n$ . The case of a linear model, (10), and of nonquadratic  $F(z)$  is the only exception. Then

$$\begin{aligned} f(x) &= \sum_{i=1}^m F_i(y_i - (u^i, x)) , & \nabla f(x) &= -\sum_{i=1}^m u^i \nabla F_i(y_i - (u^i, x)) , \\ \nabla^2 f(x) &= \sum_{i=1}^m u^i (u^i)^T \nabla^2 F_i(y_i - (u^i, x)) , \end{aligned} \quad (13)$$

and Newton's method can be applied. For nonlinear models there is a special method which uses a function of the form (12). In this method, an iteration sequence  $x^k$  is constructed and  $z(x)$  is linearized at each point, with quadratically approximated  $F(x)$ :

$$\begin{aligned}
 x^{k+1} &= \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f_k(x), \quad f_k(x) = F_k(z^k + z'(x^k)(x - x^k)), \\
 z^k &= z(x^k), \quad F_k(x) = F(z^k) + (\nabla F(z^k), z - z^k) \\
 &\quad + (\nabla^2 F(z^k)(z - z^k), z - z^k)/2.
 \end{aligned} \tag{14}$$

In particular, in the least squares method,  $F_k(z) = F(z) = \|z\|^2$ , whereas for the linear model, method (14) coincides with Newton's method. The iterative process (14) is called the Gauss-Newton method; at each step the least squares estimate is sought for the linearized model. It is possible to show that method (14) converges locally with the rate of geometric progression; the progression ratio decreases as  $z(x^k)$  decreases (i.e., the residuals at the minimum point). Also, it is possible to make method (14) converge globally, as in Newton's method (see Section 3.1).

Thus, to solve problems of the form (8), (9), one can apply either general methods—such as the conjugate gradient method, or special methods—such as the Gauss-Newton method. The former are simpler; in the latter one needs to invert the matrix at each step, but they may converge more rapidly. Numerical experiments have demonstrated that both groups of methods are almost equally efficient. Note that the accuracy of solution for problems (8), (9) is of marginal significance, since  $x^*$  does not coincide with the true value of  $x^*$ .

### 11.1.3 Robust Estimation

Maximum likelihood estimates, although asymptotically efficient, are not robust under the deviation of noise distribution law different from the one specified. Here is an example. Estimate (5) has the smallest second moment among all the unbiased estimates for normally distributed noise. Suppose however that one measurement is distributed differently, say, its variance is very large (or infinite). Then the variance of estimate (5) will also be very large or infinite. This is what actually happens in the implementation. The reason might be an instrument malfunction, a gross error in recording the data, or a computer malfunction. The probability of each occurrence is low but, nevertheless, their effect on (5) can be disastrous.

The Swiss statistician Huber suggests a technique to overcome these difficulties (1964). To begin, he assumes that the true noise density  $p$  is not known and only the class  $\mathcal{P}$  to which  $p$  belongs is known. Next, he shows the estimates which hold for all  $p \in \mathcal{P}$ . For the regression problem (10), these estimates are:

$$x_m^* = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f_m(x), \quad f_m(x) = \sum_{i=1}^m F(y_i - (u^i, x)),$$

$$F(z) = -\log p^*(z), \quad p^* = \underset{p \in \mathcal{P}}{\operatorname{argmin}} J(p), \quad (15)$$

  $J(p) = \int \frac{p'(z)^2}{p(z)} dz.$

In other words, one has to take the least favorable distribution  $p^* \in \mathcal{P}$  (with the smallest Fisher information) and then use the corresponding maximum likelihood estimate. Under natural assumptions, these estimates are asymptotically minimax on  $\mathcal{P}$ , i.e., optimal in a certain sense. We give Huber's estimates for two of the most important classes  $\mathcal{P}$ :

(i) if  $\mathcal{P}$  includes nonsingular distributions, then we have

$$x_m^* = \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f_m(x), \quad f_m(x) = \sum_{i=1}^m |y_i - (u^i, x)|, \quad (16)$$

i.e., if nothing is known about the distribution, then the least absolute value method must be used;

(ii) if  $\mathcal{P}$  is the class of “approximately normal” distributions, then we have

$$\begin{aligned} x_m^* &= \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} f_m(x), \quad f_m(x) = \sum_{i=1}^m F(y_i - (u_i, x)), \\ F(z) &= \begin{cases} z^2/2, & |z| \leq d, \\ |z|d - d^2/2, & |z| > d, \end{cases} \end{aligned} \quad (17)$$

where the parameter  $d$  depends on the “combination level” of the basic distribution. In other words, for the distribution close to normal one has to use an intermediate method of the least squares method and the least absolute value method.

Clearly, again, the robust estimation makes it necessary to solve problems of unconstrained minimization of the function (which are, however, insufficiently smooth).

One cannot use Newton's method to minimize (17). For example, if  $x$  is such that  $|y_i - (u^i, x)| > d$ ,  $i = 1, \dots, m$ , then  $\nabla^2 f_m(x) = 0$  and Newton's method is meaningless. In this case, one can use either the Levenberg-Marquardt method (Section 3.1) or first-order methods—such as the conjugate gradient method (Section 3.2). Since  $f_m(x)$  is piecewise quadratic, the conjugate gradient method can be made finite. Furthermore, there are special methods for minimizing functions of the form (17), e.g., in the Mudrov-Kushko method the functions  $F(z)$  are approximated by quadratic functions in the graph of  $F(z)$ , i.e., in each iteration the problem is solved by the least squares method:

$$\begin{aligned}
 x^{k+1} &= \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^m \mu_i^k (y_i - (u^i, x))^2, \quad z_i^k = y_i - (u^i, x^k), \\
 \mu_i^k &= \begin{cases} 1, & |z_i^k| \leq d, \\ 1/|z_i^k|, & |z_i^k| > d. \end{cases}
 \end{aligned} \tag{18}$$

Problem (16) is a problem of unconstrained minimization of a piecewise linear function. It reduces to a linear programming problem if additional variables are introduced:

$$\begin{aligned}
 \min \sum_{i=1}^m (t_i + s_i), \\
 (u^i, x) - y_i = t_i - s_i, \quad i = 1, \dots, m, \quad t_i \geq 0, \quad s_i \geq 0.
 \end{aligned} \tag{19}$$

But, in this case, the number of variables increases sharply (in regression problems, usually  $m \gg n$ ). It is convenient to apply the iterative methods of minimization of nonsmooth functions from Chapter 5, in particular, the space extension method (14) of Section 5.4. Here is one more method of solution. For the linear case we write (16) in the form

$$\min \sum_{i=1}^m |t_i|, \quad Ux - y = t$$

and form the augmented Lagrangian:

$$M(x, t, v, K) = \sum_{i=1}^m |t_i| + (v, Ux - y - t) + K \|Ux - y - t\|^2 / 2,$$

where  $v \in \mathbb{R}^m$  plays the role of the dual variables. The variable  $t$  can be eliminated since the minimum of  $M$  with respect to  $t$  is found explicitly, and the method of the augmented Lagrangian (23) of Section 9.3 becomes

$$\begin{aligned}
 x^{k+1} &= \underset{x \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^m F((u^i, x) - y_i + \frac{1}{K} v^k), \\
 v^{k+1} &= K F'((u^i, x^{k+1}) - y_i + \frac{1}{K} v^k), \\
 F(z) &= \begin{cases} z^2 / 2, & |z| \leq 1/K, \\ 1/K|z| - 1/2K^2, & |z| > 1/K. \end{cases}
 \end{aligned} \tag{20}$$

We then obtain Bertsekas's method, in which in order to solve the nonsmooth problem (16) one needs to find the minimum of a differentiable function of

the form (17) in each iteration. In other words, this is an iterative smoothing of the function (16). Since the method of the augmented Lagrangian is finite for linear programming problems (see Theorem 4 of Section 10.3), the method (20) is finite.

### Exercise

3. Find a solution of (16) for the elementary problem with  $u^i \equiv 1$ ,  $x \in \mathbb{R}^1$ .  
 ANSWER.  $x_m^*$  is the sample median, i.e., the point for which  $y_i$  contains as many larger values as there are smaller values.

#### 11.1.4 Recursive Estimation

All the estimation methods described above reduce to a complex unconstrained minimization problem. Moreover, with each new measurement this problem must be solved anew. Hence in on-line processing of data which arrives sequentially, methods such as the maximum likelihood method are inappropriate.

We shall consider a convenient, different approach to estimation. For the problem of parameter estimation in Subsection 11.1.1, let the function

$$f(x) = EQ(z, x) = \int Q(z, x) p(z, x^*) dz, \quad Q(z, x) = -\log p(z, x). \quad (21)$$

It is easy to see that  $x^*$  is the minimum point of  $f(x)$ :

$$\begin{aligned} f(x) &= -\int (\log p(z, x^*)) p(z, x^*) dz \\ &\quad - \int \log \left[ 1 + \frac{p(z, x) - p(z, x^*)}{p(z, x^*)} \right] p(z, x^*) dz \\ &\geq f(x^*) - \int [p(z, x) - p(z, x^*)] dz = f(x^*). \end{aligned}$$

Here we have used the fact that  $-\log(1+\alpha) \geq -\alpha$  for all  $\alpha$ . Thus the primal problem consists in minimizing the function (21), which is unknown since it contains the unknown density  $p(z, x^*)$ . In the maximum likelihood method the  $f(x)$  is approximated by the function (3):

$$f_m(x) = \frac{1}{m} \sum_{i=1}^m Q(z^i, x),$$

and the minimum point of this approximation is taken as  $x^*$ . One can, however, proceed differently. Let the measurements  $z^1, \dots, z^k, \dots$  arrive in sequence, and let  $x^k$  be the approximation for  $x^*$  found after  $k$  measurements have been processed. We can then assume that  $\nabla Q(z^{k+1}, x^k)$  is an approximation for  $\nabla f(x^k) = E\nabla_x Q(z, x^k)$  (cf. (1)-(3) in Section 4.1) and apply the

gradient method for minimizing  $f(x)$ :

$$x^{k+1} = x^k - \gamma_k \nabla_x Q(z^{k+1}, x^k) = x^k - \gamma_k \frac{\nabla_x p(z^{k+1}, x^k)}{p(z^{k+1}, x^k)}. \quad (22)$$

We have obtained a method of a different kind than those considered so far. Earlier,  $m$  measurements were processed simultaneously: a function of the form (3) was composed and some iterative method was applied to minimize that function. In the procedure (22), each measurement is used only once; it is not required to store all the preceding measurements, nor to work with all of the summands in (3). In such cases we speak of adaptive or recursive estimates (often, the term "stochastic approximation" is also used).

Using the results of Section 4.2, it is not hard to derive assertions on the convergence of (22) as  $k \rightarrow \infty$  under natural assumptions. In other words, estimates (22) are consistent. It might seem that these estimates are far inferior to maximum likelihood estimates (4) in terms of accuracy. But this is not the case. If we take  $\gamma_k = \gamma/k$  in (22), then for the corresponding choice of  $\gamma$  the convergence rate is of the same order as that for (4), i.e.,  $O(1/k)$ . If we consider a more general method than (22):

$$x^{k+1} = x^k - \frac{1}{k+1} J^{-1} \frac{\nabla_x p(z^{k+1}, x^k)}{p(z^{k+1}, x^k)}, \quad (23)$$

where  $J$  is the Fisher information matrix (6), then estimates (23) are asymptotically efficient. In other words, recursive estimates, although much simpler than maximum likelihood estimates, are no worse with respect to the asymptotic properties.

A similar approach is also possible for the regression problem (10). Recursive estimates of the form

$$x^{k+1} = x^k - H_k u^{k+1} \psi(u^{k+1}, x^k) - y_{k+1}) \quad (24)$$

use the measurements successively. Under general assumptions concerning the matrices  $H_k$  and the functions  $\psi: \mathbf{R}^1 \rightarrow \mathbf{R}^1$  they are consistent. Furthermore, if we take

$$\begin{aligned} H_k &= \frac{1}{k+1} J^{-1} B^{-1}, & J &= \int \frac{p'(z)^2}{p(z)} dz, \\ B &= \lim_{k \rightarrow \infty} \frac{1}{k} \sum_{i=1}^{k+1} u^i (u^i)^T, & \psi(z) &= -\frac{p'(z)}{p(z)}, \end{aligned} \quad (25) \leftarrow K$$

the recursive estimates (24) are asymptotically efficient, too. Thus, there is no need to solve complex problems like (8), and the results (asymptotically) are equivalent to estimates (8).

In the simplest case, problem (1), (2), estimates (23)-(25) assume the form  $x^{k+1} = x^k - (x^k - z^{k+1})/(k+1)$  and have the same accuracy as the nonrecursive estimate (5). In general, the recursive and nonrecursive estimates are different.

In situations where the noise distribution density  $p$  is not fully known, one can use a recursive variant of the robust estimate (15) and substitute in (24), (25) the  $-p^*(z)'/p^*(z)$  for  $\psi(z)$ , where  $p^*$  is the “least favorable” distribution of (15). The asymptotic properties remain the same as in (15).

Thus, for a large sample size recursive estimation is superior to nonrecursive estimation.

We also note that for the least squares method, a recursive scheme (see (13) in Section 4.4) can be developed, which requires no solution of systems of linear equations and gives exactly the same estimates as (11) for any number of measurements.

### Exercise

4. Show that  $J = \nabla^2 f(x^*)$  for  $f(x)$  of the form (21). Thus (23) is Newton's method (of the form (18) of Section 2.1) for minimizing  $f(x)$  in noise, but the second derivatives are not used explicitly.

#### 11.1.5 Data Analysis

We assumed in what was said that a model of a system exists and is known to within parameters. From the “outputs” of this model, which are corrupted by random errors, it was required that the unknown parameters be renewed. However, such a scheme is often too idealized. First, we do not usually know the precise form of the model; e.g., in a mathematical description of a chemical reactor secondary reactions, volumetric inhomogeneity of the process, catalytic ageing, etc. are insignificant. In other cases, an adequate model of the system is known, but it is too cumbersome to use. Lastly, in many problems the system does not have a precise mathematical description. Such are problems which involve human activity, social or economic processes, or various geological and cosmic phenomena, like detection of earthquakes, or solar activity. Furthermore, the probabilistic assumptions made above concerning the nature of noise do not frequently hold in real problems. For instance, in proving the maximum likelihood method for regression problems (8), it is necessary to assume that the noise  $\xi_i$  is independent, centered, and identically distributed with unknown density. Yet, in the implementation, the errors are, as a rule, correlated, contain a systematic component, and their distribution varies, etc. As the examples of robust estimates illustrate, all these violations of the initial assumptions are not without consequence.

All this explains why the range and validity of application of statistical identification methods are limited. Problems of describing systems,

without making assumptions regarding their probabilistic nature, are in the realm of data processing. That is, it is required to describe the available information with the aid of a model in such a way as to minimize the mismatch between the system and the model. The choice of the class of models as well as of the degree of an error is arbitrary to a certain extent, and depends on many factors, e.g., how simple the calculation of the estimates is, or how easy the resulting formulas are, among others. Of course, such empirical approach has a side effect, viz. it remains unclear how reliable the obtained description is, to what extent one can use it for purposes of extrapolation and prediction, etc. The question of choice of a model is too complicated; hence we shall examine how to match the model parameters after the model has been chosen. Here is a typical problem. Suppose the vector  $u \in \mathbf{R}^s$  is the system "input,"  $y \in \mathbf{R}^1$  is the system "output,"  $x \in \mathbf{R}^n$  are parameters,  $\phi(u, x)$  is the system model (i.e., the output prediction for fixed values of the input and parameters),  $F(z)$  is the measure of mismatch between the model and the system. We have  $u^i, y_i, i = 1, \dots, m$ , being  $m$  output measurements for the values  $u^i$  of the input variables. It is required to choose parameters such as to minimize the deviation of the predictable output from the real output, i.e., find

$$\min_{x \in \mathbf{R}^n} f(x), \quad f(x) = \sum_{i=1}^m F(y_i - \phi(u^i, x)). \quad (26)$$

Usually, the measure of mismatch is

$$F(z) = |z|, \quad (27)$$

$$F(z) = z^2. \quad (28)$$

Note that we have, again, obtained a problem of the form (12), following, however, the reasoning unrelated to mathematical statistics.

The mismatch test need not be additive, e.g., the minimax criterion is used:

$$f(x) = \max_{1 \leq i \leq m} |y_i - \phi(u^i, x)|. \quad (29)$$

It is worth noting that in the past, in the identification the quadratic function (28) was chosen as  $F(z)$  almost without exception, i.e., the least squares method. This was due to the traditional practice of extensive application of theorems of mathematical statistics concerning optimality of the least squares method, and, also, for the sake of simplicity of computations. Today, these reasons are not true any more, because the least squares method is non-robust, when the fundamental assumptions are waived and the numerical methods of minimization are developed, which allow us to solve equally easily problems with distinct  $F(z)$ .

In some cases, the information concerning the system consists of continuous, rather than discrete, measurements. For example, let  $t \in \mathbf{R}^1$  be a function of time, and let the system output  $y(t)$  be given for a known input  $u(t)$ ,  $0 \leq t \leq T$ . Then the analogs of problems (26), (27); (26), (28); (29) become

$$f(x) = \int_0^T |y(t) - \phi(u(t), x)| dt, \quad (30)$$

$$f(x) = \int_0^T (y(t) - \phi(u(t), x))^2 dt, \quad (31)$$

$$f(x) = \max_{0 \leq t \leq T} |y(t) - \phi(u(t), x)|. \quad (32)$$

Thus, we have arrived again at the problem of finite-dimensional unconstrained minimization. However, each computation of the function  $f(x)$  includes operations of integration or of taking the maximum. In particular cases, these operations are explicit. Let us assume that we solve (31) for  $\phi(u(t), x) = \sum_{i=1}^n x_i u_i(t)$ . It reduces to the problem of minimizing the function

$$f(x) = \sum_{i,j=1}^n x_i x_j \int_0^T u_i(t) u_j(t) dt - 2 \sum_{i=1}^n x_i \int_0^T y(t) u_i(t) dt. \quad (33)$$

Say, if  $u_i(t) = t^{i-1}$  (i.e., a mean square approximation of  $y(t)$  by polynomials is sought),  $T = 1$ , then we have

$$\int_0^T u_i(t) u_j(t) dt = \frac{1}{i+j-1}, \quad \int_0^T y(t) u_i(t) dt = \int_0^T y(t) t^{i-1} dt.$$

In other words, to solve this problem one needs to calculate the moments of  $y(t)$  and find the minimum of the quadratic form (33) with a Hilbert matrix

$$\left( \left[ \frac{1}{i+j-1} \right] \right)_{i,j=1}^n. \quad (34)$$

It turns out that the condition number here of the Hilbert matrix is  $\mu = 1.5 \cdot 10^7$  for  $n = 6$  and  $\mu = 1.6 \cdot 10^{13}$  for  $n = 10$ . In other words, very ill-conditioned problems arise for problems of small dimension. They have a low rate of convergence for gradient methods: according to Theorem 3 of Section 1.4 and Theorem 2 of Section 3.1, for the optimal choice of step size the gradient method converges with the rate of geometric progression, with ratio  $q \approx 1 - 2\mu^{-1}$ , i.e.,  $q \approx 1 - 10^{-7}$  for  $n = 6$  and  $q \approx 1 - 10^{-13}$  for  $n = 10$ , so that one needs to make  $\sim 10^7$  steps of the gradient method for  $n = 6$  in order to enhance the accuracy of the approximation by a factor of  $e$ . The

gradient method converges (in  $x$ ) extremely slowly already for  $n = 10$ . Moreover, the minimization problem becomes practically unstable (see Section 1.3). Indeed, errors of the order  $\varepsilon$  in determining the moments lead to errors of the order  $\mu\varepsilon$  in determining the minimum point. Hence, even if the minimum point is sought exactly but a small error ( $\sim 10^{-7}$ ) is allowed in determining the moments, it leads to an error  $\sim 1$  in determining the polynomial coefficients for  $n = 6$  and  $\sim 10^6$  for  $n = 10$ , i.e., for  $n = 10$  the resulting polynomial will be completely unrelated to the true polynomial of best mean square approximation.

However, the situation is not so dramatic. First, the given problem is criterial (see Section 6.1)—the coefficients of the hypothetical best approximation polynomial are of no consequence; it is required to define a polynomial which yields a good approximation of  $y(t)$  in the mean square sense. Hence an arbitrary vector  $x$  corresponding to values  $f(x)$  close to the minimum is acceptable. As was noted in Section 6.1, the iterative conjugate gradient-like methods converge sufficiently rapidly even for singular problems, allow one to take into account the *a priori* information concerning the solution, and are stable toward errors. As numerous experiments [11.22] demonstrate, a good variant of the conjugate gradient method yields an almost exact minimum with respect to the function. For smooth  $y(t)$  a good approximation obtains also for the solution (for  $n \leq 10$ ). Secondly, solution of the problem will be easy if we find an expansion of  $y(t)$  in orthogonal polynomials rather than the usual polynomials. Then

$$\int_0^T u_i(t) u_j(t) dt = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \end{cases}$$

hence  $f(x)$  (33) becomes

$$f(x) = \sum_{i=1}^n (x_i^2 - 2\alpha_i x_i), \quad \alpha_i = \int_0^T y(t) u_i(t) dt$$

and its minimum can be found explicitly:

$$x_i^* = \alpha_i, \quad i = 1, \dots, n.$$

Thus, no ill-posed problems arise in this case. This example shows how important it is to select the proper form of description of the model. Various changes of variables, scale transformations, etc., may significantly improve the properties of the objective function, decrease the condition number and simplify the minimization process.

## Exercises

5. Solve the following problems explicitly in  $\mathbf{R}^1$  for  $y_1 = \dots = y_{m-1} = 0, y_m = \alpha$ :

- (a)  $\min \sum_{i=1}^m (y_i - x)^2$ ;
- (b)  $\min \sum_{i=1}^m |y_i - x|$ ;
- (c)  $\min \max_{1 \leq i \leq m} |y_i - x|$ .

ANSWER: (a)  $x^* = \alpha/m$ ; (b)  $x^* = 0$ ; (c)  $x^* = \alpha/2$ . What conclusions can be drawn concerning the effect of outliers under various optimality tests?

6. It is required to approximate the given  $y_i$  at points  $t_i, i = 1, \dots, m$ , by a dependence of the form  $x_1 \exp x_2 t$ . Compare the following techniques:

- (a) find  $\min \sum_{i=1}^m (y_i - x_1 \exp x_2 t_i)^2$  for  $x \in \mathbf{R}^2$ ;
- (b) find  $\min \sum_{i=1}^m |y_i - x_1 \exp x_2 t_i|$  for  $x \in \mathbf{R}^2$ ;
- (c) make the substitution  $a_i = \log y_i, z_1 = \log x_1, z_2 = x_2$  and find  $\min \sum_{i=1}^m (a_i - z_1 - z_2 t_i)^2$  for  $z \in \mathbf{R}^2$ ;
- (d) write the differences  $\Delta y_i = y_{i+1} - y_i, \Delta t_i = t_{i+1} - t_i > 0$  and take

$$x_2^m = \frac{1}{m} \sum_{i=1}^m \frac{1}{y_i} \frac{\Delta y_i}{\Delta t_i}, \quad x_1^m = \frac{1}{m} \sum_{i=1}^m y_i \exp(-x_2^m t_i).$$

Which of the methods is simplest? Which method gives an exact solution if  $y_i = x_1^* \exp x_2^* t_i$ ? Which is more reliable (heuristically)?

### 11.1.6 Other Identification Problems

We have discussed static models of a regression system (7). In the identification of dynamic systems, one needs to deal with models described by differential, or difference, equations, and choose their parameters or initial conditions. We can take the same approaches as those for static problems, but the resulting unconstrained minimization problems have different features. For example, to calculate the gradient of the function, it is necessary to solve auxiliary linear differential large-scale equations (sensitivity equations), which can be done only in the simplest problems. Furthermore, even for linear models and quadratic tests, the objective functions may be nonquadratic with respect to the parameters sought. Or, by the physical meaning of the problem, the parameters cannot be arbitrary (e.g., the system must be stable). This entails a necessary constrained minimization. Finally, recursive estimates—such as (24)—require a special, complex, justification. Thus dynamic system identification problems are essentially more intricate than static ones.

The class of problems considered above involves processing the outcomes of “passive” tests. In some cases, it is possible to interfere in the process of data acquisition, such as the choice of input variables in the regression model (7), the choice of measuring points, and the like. This is the realm

of test design theory. There are optimization problems of three kinds.

(1) This class of problems involve the choice of an optimal test solution ( $A$ -optimal solutions,  $D$ -optimal solutions, etc.). Characteristic problems are deterministic extremum constrained problems (constraints are imposed on the input variables). The resulting problems are multimodal and they rarely permit an explicit solution. The situation becomes essentially simpler in the case of the so-called continuous solutions. In the test solution design there are many solutions of similar optimization problems. The respective optimal (or approximately optimal) solutions are organized in special catalogs.

(2) This class of problems involve processing of the test data for a fixed solution, i.e., solving the identification problem considered.

Frequently, the objective of an investigation is not to describe locally the way in which a particular index depends on other variables; instead, it is to find the extremum of this index. For example, it is required to find the multicomponent composition of maximum solidity. Using an optimal test solution, the function of relationship of the solidity to the composition in a neighborhood of an initial composition is first constructed. Next, using this dependence, a new base point is chosen, and the procedure is repeated. In this case we deal with an unconstrained minimization problem for a function, with its values calculated at each point, with random noise (see Section 4.4). It is worth noting, however, that test solution theory is separated from optimization theory and does not employ the methods of the latter. The special terminology of the test solution theory has significantly contributed to the gap between these two areas.

(3) This class of identification problems (we discuss it briefly), involve a renewal of functions in the description of the system. Problems of this kind arise in many applications, e.g., in geophysics (interpretation of electrical prospecting and seismic prospecting; inverse problems of magnetometry and gravimetry), medical sciences (cardiography and encephalography), radio engineering, among others. The identification problems of this kind lead to infinite-dimensional problems of minimization, which are, as a rule, unstable.

## Exercises

7. Consider the model  $y_{i+1} = x^*y_i + \xi_i$ ,  $y_1, \dots, y_k \in \mathbf{R}^1$  are given numbers,  $\xi_i$  are random independent kinds of noise. Write the least squares method and the maximum likelihood method for estimating the parameter  $x^* \in \mathbf{R}^1$ . Write also recursive variants of these methods.

8. A given function  $y^0(t)$  is approximately described by the equation  $dy/dt = -x^*y(t)$ , with unknown  $x^*$  and unknown initial value. What are the possible formulations of the problem of estimating  $x^*$ ? In which cases is the solution explicit? Compare the result with Exercise 6.

## 11.2 OPTIMIZATION PROBLEMS IN ENGINEERING AND ECONOMICS

### 11.2.1 Optimal Design

After a mathematical model of the designed system is stated, one can proceed to solving the optimal choice problem of parameters. Given is an index (the objective function) to be optimized. Furthermore, assume that the constraints on the admissible parameter values as well as on other features of the system are defined. The following problems are possible in such case: optimal choice of an individual unit (optimization of the characteristics of an electrical machine, an electronic instrument, a construction framework); of a technological process (optimization of a chemical reactor), of an individual manufacturing unit, or the entire industry (optimal allocation of the production line units, optimization of a gas pipeline route, or of a power network). The nonlinear objective function and nonlinear constraints of a moderate scale (less than a hundred parameters) is the typical case.

Since the characteristics of the system are usually described by complex relationships, they are frequently approximated by polynomial representations. One has then problems of the form

$$\begin{aligned} & \min f_0(x), \quad x \in \mathbf{R}^n, \\ & f_\ell(x) \leq 1, \quad \ell = 1, \dots, m, \\ & 0 < a_i \leq x_i \leq b_i, \quad i = 1, \dots, n, \end{aligned} \tag{1}$$

where

$$f_\ell(x) = \sum_{j=1}^{N_\ell} a_{ji} x_1^{b_{1j\ell}} x_2^{b_{2j\ell}} \dots x_n^{b_{nj\ell}}. \tag{2}$$

If  $a_{ji} \geq 0$  and  $b_{ij\ell}$  are arbitrary numbers, the functions (2) are called *posynomials*, and (1), (2) is called a *geometric programming problem*. Posynomials are not, in general, convex functions. Hence the general methods of nonlinear programming are ineffective for solving geometric programming problems. However, special procedures allow us to reduce problem (1), (2) to a convex problem. It is possible to make a transformation

$$z_i = \log x_i. \tag{3}$$

Then (1) becomes

$$\begin{aligned} & \min g_0(z), \\ & g_\ell(z) \leq \sum_{j=1}^{N_\ell} a_{ji} \exp \left( \sum_{i=1}^n b_{ij\ell} z_i \right), \quad g_\ell(z) \leq 1, \quad \ell = 1, \dots, m, \\ & c_i \leq z_i \leq d_i, \quad c_i = \log a_i, \quad d_i = \log b_i, \quad i = 1, \dots, n. \end{aligned} \tag{4}$$

Since  $g(z) = \exp(b, z)$  is a convex function, then (4) is a convex programming problem (with smooth  $g_\ell(z)$ ).

We shall use an unconstrained problem as an example, to show how to make problem (1) simpler. The initial problem

$$\min \sum_{j=1}^N a_j x_1^{b_{1j}} \cdots x_n^{b_{nj}}, \quad x > 0, \quad (5)$$

using the substitution (3), reduces to

$$\begin{aligned} \min \sum_{j=1}^N a_j \exp u_j, \\ (b^j, z) = u_j, \quad j = 1, \dots, N, \quad b^j = \{b_{1j}, \dots, b_{nj}\}, \end{aligned} \quad (6)$$

where the  $u_j$  are additional variables. The dual problem of (6) is:

$$\begin{aligned} L(z, u, y) &= \sum_{j=1}^N (a_j \exp u_j + y_j((b^j, z) - u_j)), \\ \psi(y) &= \inf_{z, u} L(z, u, y) = \begin{cases} \sum_{j=1}^N y_j \left(1 - \log \frac{y_j}{a_j}\right) & \text{if } \sum_{j=1}^N y_j b^j = 0, \quad y > 0, \\ -\infty & \text{otherwise,} \end{cases} \end{aligned}$$

i.e.,

$$\min \sum_{j=1}^N (y_j \log y_j - y_j(1 + \log a_j)), \quad \sum_{j=1}^N y_j b^j = 0, \quad y > 0, \quad (7)$$

and it is a problem with a convex smooth objective function and linear constraints. To solve this problem, use methods of Chapter 7: first find  $y^*$ , then calculate the  $u_j^* = \log(y_j^*/a_j)$ , solve the system  $(b^j, z) = u_j^*, j = 1, \dots, N$ , and find the optimal initial variables  $x_i^* = \exp z_i^*$ .

The following considerations are worth noting: (i) one can see again how important it is to choose appropriately the independent variables. The primal problem (1) is nonconvex and multimodal, but the substitution (3) has led to a convex problem; (ii) the considerations concerning the duality hold in the case where, at first glance, duality is inapplicable, viz. for the unconstrained minimization problem (5). Incidentally, we encountered a similar procedure before (cf. method (20) of Section 11.1 for the unconstrained minimization problem (16)).

Undoubtedly, optimal solution problems go beyond the scope of a geometric programming scheme. The impact of multimodality cannot always be so easily dispensable. A much more complicated problem arises when the design problem involves the requirement for some variables be integer-valued,

for example, parameters of a manufactured article take on only standard values; the number of machines must be an integer; the production output of the plant designed must meet the standard specifications; there is a finite number of electrical connections, or, of electronic components of a specific circuit arrangement. Combinatorial optimization problems and the methods for solving them go beyond the scope of this book, and we shall skip them.

### 11.2.2 Optimal Allocation of Resources

Problems of detecting malfunctioning, specifying the target, or designing an experiment are of a simple scheme. Suppose that we have available a particular resource. What is the optimal way to distribute this resource over  $n$  points if the utility efficiency at the  $i$ th point is given by the function  $\phi_i(x)$ ? Mathematically,

$$\begin{aligned} \min & [\phi_1(x_1) + \cdots + \phi_n(x_n)] , \\ x_1 + \cdots + x_n & = 1 , \quad x_i \geq 0, \quad i = 1, \dots, n . \end{aligned} \tag{8}$$

Here the objective function and the constraints are so simple that the solution can be obtained in explicit or “semiexplicit” form.

Let  $\phi_i(x_i)$  be convex functions (in particular, linear). Then the minimum is obtained at a vertex of the simplex defined by the constraints. There are only  $n$  vertices, and they all have the form  $\{1, 0, \dots, 0\}$ ,  $\{0, 1, 0, \dots, 0\}$ , ...,  $\{0, 0, \dots, 1\}$ . Hence it suffices to find  $j = \underset{1 \leq i \leq n}{\operatorname{argmin}} \phi_i(1)$  and take the solution to be  $x_j^* = 1$ ,  $x_i^* = 0$ ,  $i \neq j$ . In other words, the whole resource should be concentrated at the same point.

Let the  $\phi_i(x_i)$  be now convex functions. Introduce the Lagrangian

$$L(x, y) = \phi_1(x_1) + \cdots + \phi_n(x_n) + y(x_1 + \cdots + x_n - 1), \quad y \in \mathbf{R}^1 , \tag{9}$$

and the one-dimensional functions

$$\psi_i(y) = \inf_{x_i \geq 0} [\phi_i(x_i) + yx_i] . \tag{10}$$

Then

$$\psi(y) = \inf_{x \geq 0} L(x, y) = \psi_1(y) + \cdots + \psi_n(y) - y , \tag{11}$$

and the duality theorem (see Section 9.1) guarantees that the maximum point of  $\psi(y)$  be the Lagrange multiplier for the primal problem. Thus, in order to solve the problem, one needs to construct the functions  $\psi_i(y)$  in (10), find the maximum point  $y^*$  of the one-dimensional concave function  $\psi(y)$  and next find  $x_i^* = \operatorname{argmin}_{x_i \geq 0} [\phi_i(x_i) + y^*x_i]$ . This procedure can often be implemented

$\downarrow \gg$

analytically. Thus, if  $\phi_i(x_i) = (\lambda_i/2)(x_i - a_i)^2$ ,  $\lambda_i > 0$ , then

$$\begin{aligned}\psi_i(y) &= -\frac{1}{2\lambda_i}(\lambda_i a_i - y)_+^2 + \frac{\lambda_i a_i^2}{2}, \\ \psi'(y) &= \sum_{i=1}^n \frac{1}{\lambda_i}(\lambda_i a_i - y)_+ - 1,\end{aligned}\tag{12}$$

i.e.,  $\psi'(y)$  is a piecewise linear function. A root of  $\psi'(y)$  can be found by ordering the  $\lambda_i a_i$  and compute the  $\psi'(\lambda_i a_i)$  successively until the expression alternates sign, upon which  $y^*$  is found by linear interpolation.

Yet another approach to problem (8) exploits the notions of dynamic programming. Let

$$f_k(\alpha) = \min_{\substack{x_1 + \dots + x_k = \alpha \\ x_i \geq 0}} [\phi_1(x_1) + \dots + \phi_k(x_k)].\tag{13}$$

Then the  $f_k(\alpha)$  are given by the recurrence relation

$$f_{k+1}(\alpha) = \min_{0 \leq x_{k+1} \leq \alpha} [f_k(\alpha - x_{k+1}) + \phi_{k+1}(x_{k+1})],\tag{14}$$

where  $f_1(\alpha) = \phi_1(\alpha)$ , and the problem consists in calculating the  $f_n(1)$ . Thus, instead of one problem of minimizing a function of  $n$  variables we have obtained a sequence of  $n-1$  one-dimensional problems (14). Of course, each of these problems requires that one-dimensional minimization problems be solved for all parameter values  $\alpha \in [0, 1]$ . In some cases it is possible to construct the  $f_k(\alpha)$  explicitly; in other cases, approximately, by calculating the  $f_k(\alpha)$  on some network. The fact that this approach requires no assumptions concerning convexity is essential.

There are many other general problems of optimal allocation of resources (for example, large-scale resources). However, the fundamental ideas of special methods for solving those problems (using the duality theorem, or dynamic programming theorems) remain the same.

### 11.2.3 Optimal Planning

Historically, the need to solve economic problems was an impetus to the development of linear programming theory. In economics, optimal planning is the most crucial problem, and the majority of economic models are linear. Many problems of both short-term and long-term planning are reducible to linear programming problems, for an individual business firm, or a segment of the economy, optimization problems of supply and transportation, of

inventory control, etc. In economic problems, the dual variables can be treated as the resource expenditure and the production cost. The extremum conditions guarantee that the optimal plan in terms of these expenditures is also optimal for the primary problem. Thus, linear programming is more than just a tool for solving the problem; it is indeed the basis of mathematical economics theory.

Characteristic of the economic problems is the fact that they are large-scale (frequently, hundreds and even thousands of variables and constraints). The constraint matrix is usually sparse (most elements are zeros), but it can also contain thousands of nonzero elements. For numerical solution the key problem is the computer memory, as well as the computer time, and stability of computations to errors. The current sophisticated simplex-method codes is a great asset in solving these problem, but their capabilities are limited. More is expected of iterative methods (see Section 10.3). Lastly, the best help in solving large-scale problems is to use as many as possible of their specific features.

Transportation problems constitute a major class of linear programming problems:

$$\begin{aligned} & \min \sum_{i,j} c_{ij} x_{ij}, \\ & \sum_{j=1}^n x_{ij} = a_i, \quad i = 1, \dots, m, \\ & \sum_{i=1}^m x_{ij} = b_j, \quad j = 1, \dots, n, \quad x_{ij} \geq 0. \end{aligned} \tag{15}$$

We use the following notation:  $a_i$  is the production capacity at the  $i$ th point;  $b_j$  is the consumer at the  $j$ th point;  $c_{ij}$  is the unit cost of shipping the commodity from the  $i$ th point to the  $j$ th point;  $x_{ij}$  is the commodity to be transported from the  $i$ th point to the  $j$ th point. In the transportation problem, the vertices of a constraint polyhedron are easy to find (see Exercise 1). Hence there is no need for slack variables in order to find the primal vertex (see (12) of Section 10.2), and the computation in each iteration of the simplex method is essentially simple. Therefore, although in (15) the number of variables is  $m \times n$ , it is possible to solve problems with large  $m$  and  $n$  (of the order  $10^3$ ).

Many linear programming problems have a block structure:

$$\begin{aligned} & \min \sum_{i=1}^s (c^i, x_i), \\ & x_i \in Q_i, \quad \sum_{i=1}^s A_i x_i = b, \end{aligned} \tag{16}$$

where  $x_i \in \mathbb{R}^{n_i}$ ,  $Q_i = \{x_i: B_i x_i \leq d_i\}$ ,  $d_i \in \mathbb{R}^{m_i}$ ,  $A_i$  are  $m \times n_i$  matrices,  $i = 1, \dots, s$ ,  $b \in \mathbb{R}^m$ . In this problem there are  $M = m_1 + \dots + m_s + m$  con-

straints and  $N = n_1 + \dots + n_s$  variables. Problems of this kind describe objects that have weakly connected units. For example, the  $x_i$  may denote the  $i$ th manufacturer in a particular industry, and the condition  $\sum_{i=1}^s A_i x_i = b$  determines the global constraints with respect to resources and the commodity produced in all the plants together. To solve these problems, it is appropriate to use *decomposition* methods, in which the auxiliary problems are solved successively for individual units. There are many such methods. In a particular method of this kind, one constructs the Lagrange function (for the global constraints only):

$$L(x, y) = \sum_{i=1}^s (c^i, x_i) + \left[ y, \sum_{i=1}^s A_i x_i - b \right], \quad x \in \mathbf{R}^N, \quad y \in \mathbf{R}^m, \quad (17)$$

and the dual function

$$\psi(y) = \min_{\substack{x_i \in Q_i \\ i=1, \dots, s}} L(x, y) = \sum_{i=1}^s \min_{x_i \in Q_i} (c^i + A_i^T y, x_i) - (y, b). \quad (18)$$

By the duality theorem (see Section 10.1), problem (16) reduces to finding

$$\max_{y \in \mathbf{R}^m} \psi(y), \quad (19)$$

i.e., to an  $m$ -dimensional unconstrained minimization problem for a piecewise concave function. In this case the methods of Chapter 5 can be used, e.g., the subgradient method, or space-extension methods. To calculate the  $\psi(y^k)$ ,  $\partial\psi(y^k)$  one needs to solve  $s$  linear programming problems for the units:

$$x_i^k = \arg \min_{x_i \in Q_i} (c^i + A_i^T y^k, x_i) \quad (20)$$

and next (cf. (20) of Section 10.3)

$$\psi(y^k) = \sum_{i=1}^s (c^i + A_i^T y^k, x_i^k) - (y^k, b), \quad (21)$$

$$\partial\psi(y^k) = \sum_{i=1}^s A_i x_i^k - b. \quad (22)$$

A number of iterative methods of linear programming (see Section 10.3), with respect to problem (16), are also of decomposition structure and, hence, permit solution of large-scale problems. Note also that decomposition methods can be viewed not only as numerical procedures for finding the optimum, but also as mechanism of iterative planning. In particular, the foregoing method can be treated as the procedure for finding the optimal plan by a suitable adjustment of prices.

### Exercise

- L/jo*
- An arbitrary pair  $(i_0, j_0)$  is taken for problem (15). If  $a_{i_0} < b_{i_0}$  then  $x_{i_0, j_0} = a_{i_0}$  is constructed, the constraint  $\sum_{j=1}^n x_{i_0, j} = a_{i_0}$  is eliminated, and  $b_{i_0} - a_{i_0}$  is substituted for  $b_{i_0}$ . However, if  $b_{i_0} < a_{i_0}$ , then  $x_{i_0, j_0} = b_{i_0}$ , the constraint with  $b_{i_0}$  is eliminated and  $a_{i_0} - b_{i_0}$  is substituted for  $a_{i_0}$ . This procedure continues until all the constraints are eliminated. Prove that the resulting vector  $x = \{x_{ij}\}$  is a vertex of (15) and all the vertices can be defined in this way.

#### 11.2.4 Optimization Under Uncertainty

In what was said we assumed that a system and its performance have been described completely. But we do not always have such complete information; moreover, it is impossible to have complete information because there is always some element of inevitable uncertainty and randomness. For example, planning agricultural production has to account for unpredictable weather conditions; optimization of the electricity output of a generating station, or the inventory planning depends on a random demand from the consumers. Problems that involve random events are in the realm of stochastic programming. Stated in simple terms (a so-called problem of perspective stochastic programming), a decision need to be made *a priori*, before testing. This decision is an  $n$ -dimensional vector, the uncertainty is probabilistic, and the optimum as well as the constraints must be satisfied in the mean. Mathematically, the problem is

$$\begin{aligned} \min f(x), \\ f(x) &= EQ_0(x, \omega) = \int Q_0(x, \omega) dP(\omega), \quad g_i(x) \leq 0, \\ g_i(x) &= EQ_i(x, \omega) = \int Q_i(x, \omega) dP(\omega), \quad i = 1, \dots, m, \\ &x \in S \subset \mathbf{R}^n. \end{aligned} \tag{23}$$

In other words, the problem reduces to a deterministic problem in which the objective function and the constraints are mathematical expectations. The amount of information concerning a problem varies. If the functions  $Q_i(x, \omega)$  and the distribution  $P(\omega)$  are known, the problem will be equivalent, in principle, to an ordinary deterministic problem, however, each computation of the  $f(x)$  and  $g_i(x)$  and of their gradients requires the integrals be computed, and therefore is laborious. Hence it is sometimes convenient in this case to proceed as in the case of the unknown distribution  $P(\omega)$ , viz. if there is only a sample  $\omega^1, \dots, \omega^k$  of  $P(\omega)$ , then (as in the Monte-Carlo method) one can approximate  $EQ(x, \omega)$  by  $(1/k) \sum_{i=1}^k Q(x, \omega^i)$  and solve the resulting deterministic problem. Another approach (especially

useful when the realizations of  $\omega^i$  arrive in a sequence) is based on the substitution of  $Q_0(x^k, \omega^k)$  and  $\nabla Q_0(x^k, \omega^k)$  for  $f(x)$  and  $\nabla f(x)$  (similarly,  $g_i(x)$ ,  $\nabla g_i(x)$ ) at a point  $x^k$ . These methods can be regarded to be optimization methods in noise (see Chapter 4). They are referred to as methods of stochastic approximation or adaptive methods of stochastic optimization (see Section 11.1).

In stochastic programming, there are also other statements of problems, different from (23). Thus, the decision need not be made *a priori* but, rather, is made more and more precise during the observations (a so-called multistage stochastic programming problems). Or, the statement suggests that the decision per se is random, i.e., the distribution law of a random variable rather than a finite-dimensional vector is sought. For these statements of problems, one deals with infinite-dimensional optimization.

Finally, uncertainty in optimization problems does not need to lead to randomness. Thus, the objective function  $F(x, u)$  may depend on a certain parameter  $u$ , which is not random but is still unknown (it is known only that  $u \in U$ , where  $U$  is some subset of  $\mathbf{R}^m$ ). In this case, the optimization problem can be stated in various ways. The minimax approach is most common, i.e., anticipation of the worst value of the parameter. Then the problem becomes

$$\begin{aligned} & \min f(x), \\ & f(x) = \max_{u \in U} F(x, u) \end{aligned} \tag{24}$$

(for the sake of simplicity, we assume that there are no constraints on  $x \in \mathbf{R}^n$ . In particular, if  $u$  can take on only a finite number of values, we obtain the simplest minimax problem:

$$\begin{aligned} & \min f(x), \\ & f(x) = \max_{1 \leq i \leq m} F_i(x). \end{aligned} \tag{25}$$

The approach to solving minimax problems is twofold.

1) (24) can be viewed as the problem of unconstrained minimization of a nonsmooth function ( $f(x)$  of the form (25) is generally nondifferentiable, even if the  $F_i(x)$  are smooth; say, if the  $F_i(x)$  are affine:  $F_i(x) = (a^i, x) - b_i$ , then  $f(x)$  is piecewise linear, i.e., a nonsmooth function). In this case, if the  $F_i(x)$  are convex in  $x$ , then, using the rule for determining the subgradient (Lemma 11 in Section 5.1), it is easy to find the subgradient of  $f(x)$  of the form (25). Thus, to minimize  $f(x)$ , one can use the methods of Chapter 5. For functions of the form (24), the  $\varepsilon$ -subgradient is determined (Lemma 13 of Section 5.1), and there is no need to seek the maximum for  $u \in U$  exactly. Hence the  $\varepsilon$ -subgradient method (11) of Section 5.3 is appropriate.

2) (25) can be reduced to a nonlinear programming problem by means of the slack variable  $t \in \mathbf{R}^1$ :

$$\begin{aligned} & \min t, \\ & F_i(x) \leq t, \quad i = 1, \dots, m. \end{aligned} \tag{26}$$

Problem (26) is special since Slater's condition is satisfied, the choice of an admissible point is easy (it suffices to take an arbitrary  $x_0$  and  $t^0 \geq \max_{1 \leq i \leq m} f_i(x^0)$ ), or, the variable  $t$  is eliminated from the auxiliary problems for most of the methods. To illustrate, we use the methods based on the Lagrangian. Let

$$L(x, t, y) = t + \sum_{i=1}^m y_i(F_i(x) - t). \tag{27}$$

It is seen that in  $\min_{t \in \mathbf{R}^1} L(x, t, y) = -\infty$  if  $\sum_{i=1}^m y_i \neq 1$ . Hence, method (16) of Section 9.3 becomes

$$\begin{aligned} x^{k+1} &= \underset{x}{\operatorname{argmin}} L(x, y^k), & L(x, y) &= \sum_{i=1}^m y_i F_i(x), \\ y^{k+1} &= P_S(y^k + \gamma F(x^{k+1})), & S &= \left\{ y : y \geq 0, \sum_{i=1}^m y_i = 1 \right\}, \end{aligned} \tag{28}$$

$$F(x) = \{F_1(x), \dots, F_m(x)\}.$$

Other nonlinear programming methods can be similarly modified with respect to problem (26), as well.

## Exercises

2. Write extremum conditions for problem (25) for convex  $F_i(x)$ :
  - (a) from the condition  $0 \in \partial f(x^*)$  (see Section 5.2) and
  - (b) from the Kuhn-Tucker theorem for problem (26).
3. In problem (25), formulate conditions under which a sharp minimum is attained (see (9) of Section 5.2 and (11) of Section 9.2).

### 11.2.5 Extremal Control

The theory of extremal systems has emerged from the classical theory of automatic control, in particular from equilibrium point problems. Under investigation is the problem described by a certain performance index, such as an efficiency coefficient (for electric power generating stations), or machinery productivity (for technological processes), or the net fallout (for chemical reactors). The performance index depends on the number of

system parameters. Control is used to select adjustable parameters so as to determine the extremal value of the performance index. Characteristic of the problem is the fact it has no analytic description of the relationship of the performance index and the parameters. Thus, extremal control theory is related to real-time optimization of operating systems, an adequate model of which is unknown. Due to the presence of noise, the system output is a random variable.

Mathematically, these problems can be regarded to be problems of minimizing a function concerning which the only information we have consists of the values calculated with random errors. Methods for solving such problems were given in Section 4.4. It is worth noting that extremal control problems have specific features: (i) since optimization is performed at an operating system, it is essential to obtain small values of the function at the end of the computations as well as during the entire iterative process. Hence, in particular, large steps may lead to an increase of the function and therefore must be avoided; (ii) the controlled system is, as a rule, non-stationary. Since the fluctuations of independent parameters and the drift of the characteristics are unavoidable, the objective function changes in time, i.e., we deal with the conditions described in Section 6.3; and (iii) optimization is executed in continuous (not discrete) time, and the controlled system has a time delay. All these factors contribute to the situation that extremal control techniques in the practical implementation differ appreciably from the standard algorithms of unconstrained minimization.

### 11.2.6 Optimal Control

Optimal control problems constitute another major class of extremal problems based on automatic control problems. The problem consists in finding a law of variation of control variables for a dynamic system, as to minimize a certain optimality test. The system is assumed to be described by ordinary differential equations, which are known (as well as the optimality test). Optimal control is widely used, for example, in problems of selecting the orbit for an aircraft or a spacecraft, operation modes for electric generators, or for chemical reactors. The greatest contributions to optimal control theory are due to P.L. Pontryagin (Pontryagin's Maximum Principle) and R. Bellman (Bellman's Optimality Principle). Other important results pertain to numerical methods.

We shall describe a discrete variant of the optimal control problem:

$$\min \left[ F(x_N) + \sum_{i=0}^{N-1} f_i(x_i, u_i) \right],$$

$$\begin{aligned}
 x_{i+1} &= \phi_i(x_i, u_i), & i = 0, 1, \dots, N-1, \quad x_0 = a, \\
 u_i &\in U_i, & i = 0, 1, \dots, N-1, \\
 x_i &\in X_i, & i = 1, \dots, N.
 \end{aligned} \tag{29}$$

Here  $x_i \in \mathbf{R}^n$ ,  $i = 0, \dots, N$ , are the states of the process or the phase variables,  $u_i \in \mathbf{R}^m$ ,  $i = 0, \dots, N-1$ , are the controls or control variables, the relations  $x_{i+1} = \phi_i(x_i, u_i)$  are the state equations and determine the trajectory of the process  $\{x_0, x_1, \dots, x_N\}$ ,  $u_i \in U_i$  and  $x_i \in X_i$ , are the constraints on the controls and states, respectively. Special cases of problem (29) are the problem of optimization of the terminal state (the terminal problem) without constraints on the states:

$$\begin{aligned}
 \min F(x_N), \\
 x_{i+1} &= \phi_i(x_i, u_i), & i = 0, 1, \dots, N-1, \quad x_0 = a, \\
 u_i &\in U_i, & i = 0, \dots, N-1,
 \end{aligned} \tag{30}$$

or the discrete approximation of the classical problem of the calculus of variations:

$$\begin{aligned}
 \min \sum_{i=0}^{N-1} F_i(x_i, u_i), \\
 x_{i+1} &= x_i + \varepsilon u_i, & x_i \in \mathbf{R}^1, \quad u_i \in \mathbf{R}^1, \quad i = 0, \dots, N-1, \\
 x_0 &= a, \quad x_N = b,
 \end{aligned} \tag{31}$$

among others. The approach to problem (29) is twofold: 1) treat the vectors  $x_i$ ,  $u_i$  as independent variables, and the state equations as equality constraints. Such a straightforward approach is not very effective since we obtain the problem with too many variables and constraints, while the specific structure of problem (29) is practically ignored, and 2) one can treat only  $u_i$  as independent variables and use them to express  $x_i$  by the state equations and the initial condition for  $x_0$ . Thus, problem (3) can be written as

$$\begin{aligned}
 \min f(u) \\
 u &= \{u_0, \dots, u_{N-1}\} \in \mathbf{R}^{Nm}, \\
 u &\in U, \quad U = U_0 \times \dots \times U_{N-1},
 \end{aligned} \tag{32}$$

where  $f(u) = F(x_N)$ , and  $x_N = x_N(u)$  is found recursively from the relations  $x_{i+1} = \phi_i(x_i, u_i)$ ,  $i = 0, \dots, N-1$ ,  $x_0 = a$ .

Suppose the functions  $F(x_N)$ ,  $\phi_i(x_i, u_i)$  are differentiable in the arguments. Let us calculate the gradient of  $f(u)$ . Let  $\bar{u} = \{\bar{u}_0, \dots, \bar{u}_{N-1}\}$  denote the control increment. Then  $\bar{x}$ , the linear part of the state increment,

is described by

$$\begin{aligned}\bar{x}_{i+1} &= \phi'_x(x_i, u_i)\bar{x}_i + \phi'_u(x_i, u_i)\bar{u}_i, \quad i = 0, \dots, N-1, \\ \bar{x}_0 &= 0.\end{aligned}$$

Introduce the adjoint system

$$\begin{aligned}p_i &= \phi'_x(x_i, u_i)^T p_{i+1}, \quad p_i \in \mathbf{R}^n, \quad i = 0, \dots, N-1, \\ p_N &= \nabla F(x_N).\end{aligned}$$

Then

$$\begin{aligned}(p_{i+1}, \bar{x}_{i+1}) &= (p_{i+1}, \phi'_x(x_i, u_i)\bar{x}_i + \phi'_u(x_i, u_i)\bar{u}_i) \\ &= (p_i, \bar{x}_i) + (\phi'_u(x_i, u_i)^T p_{i+1}, \bar{u}_i).\end{aligned}$$

Summing these inequalities yields

$$(\nabla F(x_N), \bar{x}_N) = \sum_{i=0}^{N-1} (\phi'_u(x_i, u_i)^T p_{i+1}, \bar{u}_i)$$

and

$$\nabla f(u) = \{\phi'_u(x_0, u_0)^T p_1, \dots, \phi'_u(x_{N-1}, u_{N-1})^T p_N\}. \quad (33)$$

Using the necessary minimum conditions for  $f(u)$  on  $U$  (Theorem 1 of Section 7.1), we find that if the  $U_0, \dots, U_{N-1}$  are convex, the control  $u^*$  is optimal for (30) and  $x^*$  is the corresponding trajectory, then the condition (local maximum principle) is satisfied:

$$(\phi'_u(x_i^*, u_i^*)^T p_{i+1}^*, u_i - u_i^*) \geq 0 \quad \forall u_i \in U_i, \quad (34)$$

where  $p_i^*$  is the solution of the adjoint system

$$p_i^* = \phi'_x(x_i^*, u_i^*)^T p_{i+1}^*, \quad i = 0, \dots, N-1, \quad p_N^* = \nabla F(x_N^*). \quad (35)$$

Using  $\nabla f(u)$  in the form (33), it is not hard to write the iterative methods for solving problem (32), described in Section 7.2. Thus, in the gradient projection method the sequence of approximations  $u^k$  is constructed as follows. First we determine the trajectory  $x^k$  from the state equations

$$x_{i+1}^k = \phi_i(x_i^k, u_i^k), \quad i = 0, \dots, N-1, \quad x_0^k = a,$$

then compute recursively (from  $N$  to 0)

$$p_i^k = \phi'_x(x_i^k, u_i^k)^T p_{i+1}^k, \quad i = N-1, \dots, 0, \quad p_N^k = \nabla F(x_N^k),$$

and finally determine  $u^{k+1}$ :

$$u_i^{k+1} = P_{U_i}(u_i^k - \gamma \phi'_u(x_i^k, u_i^k)^T p_{i+1}^k).$$

This approach to problem (30) is simple and makes efficient use of its special features. However, the phase constraints make the situation difficult, and then we need to do the following: Assume that in problem (29) there are no constraints on the control and the equation  $x_{i+1} = \phi_i(x_i, u_i)$  has a unique solution with respect to  $u_i \in \mathbf{R}^m$  for any  $x_i, x_{i+1} \in \mathbf{R}^n$ . Note that this situation is not typical, since usually  $m < n$ ; but it is exactly the case in problem (31). Then, we eliminate the controls by means of the state equations and obtain the problem

$$\min \sum_{i=0}^{N-1} \Phi_i(x_i, x_{i+1}), \quad (36)$$

$$x_i \in X_i, \quad i = 1, \dots, N, \quad x_0 = a.$$

To solve (36), we can apply either the general gradient projection methods (they are simple because of the constraints in (36)), or special methods, e.g., the dynamic programming method, in which the functions  $V_i(x_i)$  are given by the recurrence relations:

$$V_{i+1}(x_{i+1}) = \min_{x_i \in X_i} [\Phi_i(x_i, x_{i+1}) + V_i(x_i)], \quad i = 1, \dots, N-1, \quad (37)$$

$$V_0(x_0) = \Phi_0(a, x_1),$$

the minimum of  $V_N(x_N)$  with respect to  $x_N \in X_N$  yielding the solution (cf. (14)). On the other hand, to minimize (36), it is convenient to apply the coordinatewise descent methods (i.e., minimize successively in  $x_i \in X_i$ ,  $i = 1, \dots, N$ , for fixed values of the remaining variables). Indeed, in (36) only the values of the functions  $\Phi_{i-1}(x_{i-1}, x_i)$  and  $\Phi_i(x_i, x_{i+1})$  are calculated, and the constraints on each variable  $x_i \in X_i$  are determined independently. Such a method is called a method of local variations.

## Exercises

4. Write the resource allocation problem (8) in terms of an optimal control problem.
5. Show that if  $\phi_i(x_i, u_i) = A_i x_i + B_i u_i$ ,  $F(x_N)$  is convex and the  $U_i$  are convex, then (34) is a sufficient extremum condition.
6. Find an explicit form of solution of problem (31), where  $F_i(x_i, u_i) = \alpha_i x_i^2 + \beta_i u_i^2$ ,  $\alpha_i \geq 0$ ,  $\beta_i \geq 0$ : (i) from extremum conditions; (ii) through the dynamic programming method.

7. What are the possible ways of solving the linear dynamic programming problem:  $\min (c_N, x_N)$ ,  $x_{i+1} = A_i x_i + B_i u_i$ ,  $x_i \in X_i$ ,  $u_i \in U_i$ ,  $x_0 = a$ , where  $X_i$ ,  $U_i$  are polyhedral sets.

### 11.3 OPTIMIZATION PROBLEMS IN MATHEMATICS AND PHYSICS

Extremum problems were treated extensively in mathematics in the past, long before a general theory of optimization was born. Problems pertaining to variational principles were examined in mechanics, optics and other areas of physics. Here are a few examples.

#### 11.3.1 Optimal Approximation Problems

Given is an overdetermined system of linear equations

$$(a^i, x) = b_i, \quad i = 1, \dots, m, \quad x \in \mathbb{R}^n. \quad (1)$$

Such a system may not have a solution (especially for  $m > n$ ). Therefore, it is appropriate to pose the problem of minimizing some norm of a residual. In particular,

$$\min \sum_{i=1}^m |(a^i, x) - b_i|, \quad (2)$$

$$\min \sum_{i=1}^m ((a^i, x) - b_i)^2, \quad (3)$$

$$\min \max_{1 \leq i \leq m} |(a^i, x) - b_i| \quad (4)$$

are respectively the problem of best approximation in the norm  $\ell_1$ , best mean square approximation, and best uniform (or Chebyshev) approximation. Note that such problems arise also in applied identification problems (see Section 11.1). Next we show how the general theory can be applied to analyze problems (2)-(4). Using Theorem 9 of Section 10.1 and Theorem 4 of Section 10.4, we obtain that these problems always have solutions. A normal solution (see Section 6.1) of problem (3) can be written in the explicit form:

$$x^* = A^+ b, \quad (5)$$

where  $A$  is an  $m \times n$  matrix with rows  $a^i$ ,  $A^+$  is the pseudoinverse,  $b = (b_1, \dots, b_m)$ . Now we derive extremum conditions in (4). Using Lemma 11 of Section 5.1 and Theorem 1 of Section 5.2, we convince ourselves that a point  $x^*$ ,  $f(x^*) > 0$ ,  $f(x) = \max_{1 \leq i \leq m} |(a^i, x) - b_i|$ , is a solution of (4) iff 0 belongs

to the convex hull of the vectors  $\alpha_i a^i$ ,  $i \in I^*$ , where  $I^* = \{i: |(a^i, x^*) - \beta_i| = f(x^*)\}$ ,  $\alpha_i = \text{sign } ((a^i, x^*) - b_i)$ . Then, by Caratheodory's lemma (Lemma 1 of Section 5.1) 0 is a convex combination of  $n+1$  vectors. Thus, a necessary and sufficient extremum condition is: there are  $n+1$  integers  $\lambda_i \geq 0$ ,  $i \in I^*$ , such that

$$\sum_{i \in I^*} \lambda_i \alpha_i a^i = 0, \quad \sum_{i \in I^*} \lambda_i = 1. \quad (6)$$

Along with the discrete problems (2)-(4) it is possible to consider also their continuous analogs. Let a function  $b(t)$  and a system of functions  $a_1(t), \dots, a_m(t)$  be given on an interval  $[0, 1]$ . Problems of best approximation of  $b(t)$  for the given system in the norms  $L_1, L_2, L_\infty$ , respectively, are

$$\min \int_0^1 \left| b(t) - \sum_{i=1}^m x_i a_i(t) \right| dt, \quad (7)$$

$$\min \int_0^1 \left( b(t) - \sum_{i=1}^m x_i a_i(t) \right)^2 dt, \quad (8)$$

$$\min \max_{0 \leq t \leq 1} \left| b(t) - \sum_{i=1}^m x_i a_i(t) \right|. \quad (9)$$

Here we cannot obtain results concerning the existence of a solution and the extremum conditions as before, simply and without any additional assumptions on  $b(t)$  and  $a_i(t)$ . Yet, we can make an analysis of this problem. In particular, extending Lemma 11 of Section 5.1 to the case where the number of functions (of the form  $f(x) = \max_{0 \leq t \leq 1} f(x, t)$ ) is finite, we can derive the classical Chebyshev theorem on best uniform approximation for the case where  $a_i(t)$  is a polynomial of degree  $i-1$ .

We dwelled briefly on the simplest best approximation problems. Our objective was to show the way in which approximation theory is related to optimization theory. Of course, approximation problems have special features of their own, and there is a very sophisticated mathematical apparatus to analyze them. Nevertheless this relationship is useful in examining best approximation conditions and especially in developing numerical methods for solving approximation problems.

### Exercise

1. Consider the discrete problem of best approximation by polynomials, i.e., problem (4) with  $a^i = \{1, t_i, \dots, t_i^{n-1}\}$ ,  $0 \leq t_1 \leq \dots \leq t_m \leq 1$ . Derive

a discrete analog of Chebyshev's theorem: for  $x^*$  to be a solution it is necessary and sufficient that the quantities  $|\varepsilon_i|$  (where  $\varepsilon_i = \sum_{j=1}^n x_j^* t_i^{j-1} - b_i$ ) attain a maximum at no less than  $n+1$  points and the signs of the  $\varepsilon_i$  alternate at these points. *Hint:* Treat (6) as a system of equations in the  $n+1$  variables  $z_i = \lambda_i \alpha_i$ . Find the solution and use the property that the Vandermonde determinant changes sign whenever two columns are interchanged.

### 11.3.2 Geometric Extremum Problems

In geometry, there arise many maximum and minimum problems with differing formulations and solutions.

One class of problems involve finding bodies (specified to within parameters) which have a minimum volume, surface or other similar characteristics. We dealt with one such problem in constructing the ellipsoid method (see (12) of Section 5.4), in which it was required to find an ellipsoid of smaller volume, circumscribed around a circle. The problem is easily reduced to that of minimizing a function of one variable and can be solved explicitly.

Another class of problems are those in which the form of the set with an extremal property is not fixed, e.g., find the radius of the smallest ball into which any circle of diameter 1 can be inscribed (Young's theorem gives the solution); or, find the radius of the smallest ball which can be placed inside an arbitrary convex set of width 1 (Blaschke's theorem gives the solution). An example of solution of an extremal problem of this kind is provided by Lemma 1 of Section 5.4, used in proving the center-of-gravity method.

The same class of problems include the so-called isoperimetric problems, in which it is required to optimize one of the geometric characteristics of a closed set (say, the volume) when values of other characteristics (say, the surface) are fixed. Such is Dido's problem of finding the curve, with a given parameter, which encloses the maximum area in  $\mathbb{R}^2$ , formed by an arc of length 1 with endpoints on the line. Similar problems are examined in the calculus of variations and are infinite-dimensional.

Also, there are many problems like the problem of maximum density packing of objects, or, the problem of finding the minimal  $\varepsilon$ -network, among others. They are of a combinatorial nature, and usually very complex.

We are not going into detailed discussion of all these problems. To illustrate how optimization theory can be fruitfully used in analyzing geometric problems, we give two examples.

Helly's theorem (this problem is in no way connected with optimization):

If  $A_i$ ,  $i = 1, \dots, m$ , are convex sets in  $\mathbb{R}^n$ ,  $A_1$  is bounded and the intersection of any  $n+1$  sets in  $\{A_i\}$  is nonempty, then all the sets have a common

point. Introduce the function  $f(x) = \max_{1 \leq i \leq m} \rho(x, A_i)$ , where  $\rho(x, A_i)$  is the distance from  $x$  to  $A_i$ . The functions  $\rho(x, A_i)$  and  $f(x)$  are convex (see Exercise 2 of Section 5.1); the set  $\{x: f(x) \leq \alpha\} \subset \{x: \rho(x, A_1) \leq \alpha\}$ , and therefore is bounded for any  $\alpha \geq 0$ . Hence (Theorem 3 of Section 5.3),  $f(x)$  attains a minimum on  $\mathbf{R}^n$  at some point  $x^*$ , and therefore (Theorem 1 of Section 5.2)  $0 \in \partial f(x^*)$ . Using the subgradient of  $f(x)$  (Lemma 11 of Section 5.1), as well as Caratheodory's lemma (Lemma 1 of Section 5.1), we obtain that there are  $n+1$  indices (for the sake of definiteness, the first  $n+1$ ) such that  $\rho(x^*, A_i) = f(x^*)$ , and numbers  $\lambda_i \geq 0$ ,  $i = 1, \dots, n+1$ ,  $\sum_{i=1}^{n+1} \lambda_i = 1$ , such that  $\sum_{i=1}^{n+1} \lambda_i \partial \rho(x^*, A_i) = 0$ . But this is a sufficient condition of unconstrained minimization (Theorem 1 of Section 5.2) for the function  $\phi(x) = \max_{1 \leq i \leq n+1} \rho(x, A_i)$ . Since  $\bigcap_{i=1}^{n+1} A_i \neq \emptyset$  (by hypothesis), then  $\min_{x \in \mathbf{R}^n} \phi(x) = 0$ , i.e.,  $\phi(x^*) = 0$ ,  $\rho(x^*, A_i) = 0$ ,  $i = 1, \dots, n+1$ . But  $\rho(x^*, A_i) = f(x^*)$ ,  $i = 1, \dots, n+1$ , and hence  $f(x^*) = 0$ , which does indeed imply that  $x^* \in A_i$ ,  $i = 1, \dots, m$ .

Of course, Helly's theorem can be proved by purely geometric means, but our proof is still instructive since it shows that it is possible to apply standard techniques to solve nonstandard problems.

### Steiner's problem:

Find a point in  $\mathbf{R}^n$  for which the sum of distances to  $m$  fixed points  $a^1, \dots, a^m$  is minimal. In other words, we seek

$$\begin{aligned} & \min f(x), \\ & f(x) = \sum_{i=1}^m \|x - a^i\|. \end{aligned} \tag{10}$$

The fact that the minimum is attainable and unique (if only the  $a^i$  are not collinear) has been observed in Exercises 4 and 5 of Section 5.2. If the minimum point  $x^*$  does not coincide with one of the points  $a^i$ , then  $f(x)$  is differentiable at  $x^*$  (see Exercise 1 of Section 1.1) and (Theorem 1 of Section 1.2)

$$\nabla f(x^*) = \sum_{i=1}^m e^i = 0, \quad e^i = \frac{x^* - a^i}{\|x^* - a^i\|}. \tag{11}$$

In some cases this equation has a solution. Suppose, for instance, that  $n = 2$ ,  $m = 3$ , i.e., we seek the closest point to the vertices of a triangle. It follows from (11) that  $e^1 + e^2 + e^3 = 0$ ,  $\|e^i\| = 1$ , and this is possible iff the vectors  $e^i$  form equal angles with each other. Thus, the point  $x^*$  is such that the segments joining it with the vertices form a  $120^\circ$  angle (such a point is called a Torricelli point, see Fig. 43(a)). If there is no such a

point (e.g., if one of the angles in the triangle is not less than  $120^\circ$ ), then the minimum of  $f(x)$  obtains at the respective vertex (Fig. 43(b)). In the general case, to minimize (10), one has to use numerical methods. In a neighborhood of the minimum the  $f(x)$  can be either smooth (if  $x^*$  does not coincide with  $a^1, \dots, a^m$ ) or nonsmooth (if  $x^* = a^i$ ). Hence it is appropriate to combine smooth minimization methods with nonsmooth ones.

### Exercises

2. Let  $A_1, \dots, A_m$  be closed convex sets in  $\mathbf{R}^n$  and let  $A = A_1 \cap \dots \cap A_m \neq \emptyset$ . Prove that the method of successive projections of the form  $x^{k+1} = P_k(x^k)$ , where  $P_k$  is the projection operator onto the set most distant from  $x^k$ , converges to some point in  $A$ . Hint: Treat this method as a subgradient method of the form (7) in Section 5.3, for minimizing  $f(x) = \min_{1 \leq i \leq m} \rho(x, A_i)$ .
3. Prove that the shortest path joining the vertices of a square has the form shown in Fig. 43(c).

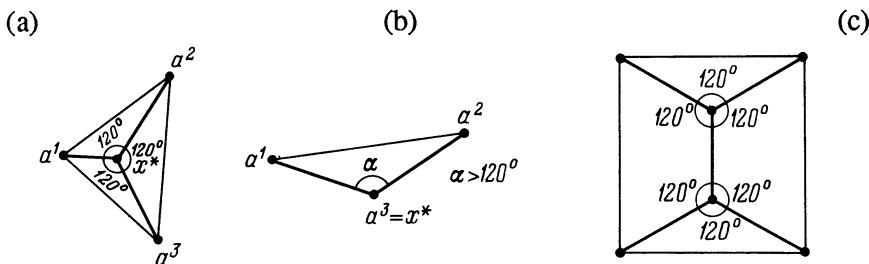


Fig. 43 Steiner's problem.

### 11.3.3 Variational Principles in Physics

The problems of Section 11.2 lead to optimization problems because it is possible to choose a particular decision and a control. It is surprising that many processes and phenomena of Nature, although not controllable, can be still described by means of variational principles—such as the variational principles of mechanics (the principle of least action, the principle of minimum potential energy), of optics (Fermat's principle of least time), of hydraulics, etc. Processes of finding the equilibrium in such systems can be viewed as particular methods for solving extremal problems applied by Nature itself. In the earliest days of mathematical programming attempts were made to utilize directly electrical, hydraulic, and other similar means in solving optimization problems. Today, in the era of digital computers, these physical

models are not needed any more. However, an analysis of physical analogies can be quite useful in designing numerical methods of optimization. In the heavy-ball method in Section 3.12, the model (motion of a body in a potential field in the presence of friction force) enabled us to evaluate qualitatively the behavior of an iterative process (possible acceleration through inertia, the ability to “skip” small local minima, and so on). Of course, our inferences need a rigorous mathematical justification (like Theorem 1 of Section 3.2), but their heuristic value is unquestionable.

On the other hand, some results of mathematical optimization theory proved to be of value in studying physical problems. Thus, variants of the duality theorem (conceptually) have been known in analytic mechanics and the theory of electric circuits; the comprehension of the mathematical nature of these assertions contributed to a better understanding of the methodology and helped extend these assertions to a larger class of systems.

Most widely used are numerical optimization methods for solving physical problems, e.g., the problem of determining statistically indeterminate elastic systems, or variational problems of fluid medium equilibrium, or the problem of computing elasticity or viscosity friction deformations (non-smooth).

### Exercise

4. Light is propagated in one medium with velocity  $v_1$ , in another with velocity  $v_2$ , and the boundary separating the media is assumed to be rectilinear. Derive from Fermat's principle (light is propagated between two points along the path requiring the least time) Snell's law of refraction:  $(\sin \alpha_1)/(\sin \alpha_2) = v_1/v_2$ , where  $\alpha_1$  is the angle of incidence,  $\alpha_2$  is the angle of refraction.

## **CHAPTER 12**

### **OPTIMIZATION PROBLEMS: IMPLEMENTATION**

In this chapter we shall discuss the practical implementation of optimization problems, including the methodology of problem formulation, criteria for choosing a method, and numerical solution and analysis of the solutions obtained. Also, we examine the currently available software and some numerical results for solving test optimization problems.

#### **12.1 SOLUTION OF A PROBLEM**

We make no claim that our methodological remarks concerning the statement and solution of optimization problems are applicable to all. The variety of problems (even most fundamental problems are too numerous) makes it an impossible task to cover them all. Also, to know how to solve a practical problem requires great talent and skill, and this knowledge can be acquired only by trial and error. Finally, the author may have his own, subjective attitudes and opinions.

##### **12.1.1 The Mathematical “Formalization” of a Problem**

Rarely does an optimization problem in practice come with a ready-made mathematical formulation. As a rule, the first step in solving the problem is to state it in mathematical terms and formalize it. At this stage, the “customer,” an expert in the application area, must cooperate with the “analyst,” a mathematician. It is best when these specialists are combined in one; otherwise they are likely to have difficulty communicating with each other.

The analyst has to understand and appreciate the practical nuances of the problem. The user must clearly understand the objective of the investigation, the application of the resulting solution, the accessibility and accuracy of the primary data, the relative importance of the various assumptions, and so on. Sometimes, even at this early stage, it may turn out that the problem as stated does not have a solution, (e.g., due to lack of data), or is pointless (because the optimization payoff is null).

At the next step, the optimality criterion as well as constraints need to be formulated in theoretical terms. In real world problems one often encounters the multicriteria phenomenon; for example, in designing a new unit it is desirable that all of its characteristics should be optimal (i.e., productivity, cost, weight, reliability). Usually, it is hard to achieve this all at once. Then, the approach is twofold: either devise a unified general criterion or specify the feasible values of all the characteristics of the unit except one characteristic, and optimize it.

The next step is the mathematical formalization of the problem. The experience and know-how of the analyst is of great importance at this stage, to provide the necessary mathematical sophistication, as well as specialized skills in the numerical solution of a concrete problem. At this particular stage one needs to choose independent variables and write the objective function and the constraints. The user has to indicate which of the parameters and conditions are essential and which may be ignored; under which conditions it is okay to linearize the functions; which conceptual constraints must be taken into account. At the same time, the analyst constructs the simplest model (if possible, a linear model with a minimum number of variables and constraints). The result is a preliminary mathematical “formalization” of the problem.

This formalization must be carefully checked, in order to see whether each element has been taken into account, or, whether it is too crude. Now is the time for the preliminary mathematical investigation of the problem, i.e., define the class the problem belongs to, has it ever been solved before, and if so, how successfully, which hardware and software are currently available, etc. One should also evaluate the general features of the problem, say, whether it is convex or not, how smooth the functions are in the formulation of the problem, wherether there is any risk of having multimodality, how great the computational errors are, whether the problem is strongly degenerate, among others. In this respect, it is frequently necessary to modify the statement of the problem, for example, choose new variables in which the objective function is better conditioned or is convex, write one nonsmooth constraint as several smooth ones, or go over to a dual problem with fewer variables, etc.

In the model constructed, not all of the quantities are, as a rule, known. The problem of acquiring the needed initial data is often one of the most difficult. This is especially true for economic problems, in which the acquisition of reliable and sufficient statistical data is extremely problem-

atic. In engineering problems one needs to solve auxiliary identification parameter problems; the accuracy (at least, in qualitative terms) of the results obtained is very essential. At the end of this elaborate and multi-stage procedure is the final mathematical formulation of the problem to be optimized.

### 12.1.2 The Choice of Methods and Codes

Only in exceptional cases, as was illustrated above, can the solution of an optimization problem be found explicitly. Usually, one has to seek the extremum numerically, using a computer. In choosing the method and the computer program one needs first to take into consideration the computer capabilities, the available software, and the total number of similar problems to be solved.

If a single optimization problem is being solved and it is not too complicated, then finding the best method and code is hardly worth the time and effort. It is enough that the software be usable for problems of similar type and size. The fact is that the computer time for solving the problem is greater than that for a specially written program (for example, one hour instead of one minute) is not too essential. Similarly, if there is no software for a particular problem, it is only natural to write the program for the simplest method for solving this problem, and ignore the rate of convergence and other characteristics of the method.

Only when similar problems have to be solved in bulk (this is typical for optimization problems) the question of choosing a numerical method becomes crucial, due to the increasing loss of the computing time or of the accuracy of solution. In this case, it makes sense to test differing existing computer routines on a few examples: if the results prove to be unsatisfactory (the convergence is too slow or the desirable accuracy of solution is unobtainable), a special routine is needed for solving the given class of problems. In selecting the method, the following factors are essential: the rate of convergence, the volume of computations in each iteration, stability under noise of different kinds, the needed memory (all these factors have been examined in the earlier chapters). After the method has been selected on the basis of the *a priori* arguments, it is appropriate to turn to its algorithmic and programming implementation. It is not easy to develop a well-defined algorithm relying on the level of detailing the methods which we are using in this book. We need to specify the subroutines for solving the auxiliary problems, define the parameters of the accuracy of their solution, determine the stopping criteria, etc. They determine, to a great extent, the degree of effectiveness of the algorithm. Similar problems arise in writing the program. The language of the program, data files, methods of realization of arithmetic expressions are of major importance. We noted before that, in particular, the refinement of all such details in the simplex-

method software has contributed enormously to successful performance of finite linear programming methods in problem-solving practice.

After the first variant of the routine is programmed and debugged, the algorithm is ready for fine tuning. Several standard problems are then tested. Usually the test fails; one must then find out why it failed and try to correct it. To do this, we can vary the algorithm parameters, replace individual subroutines (say, the one-dimensional search procedure), make check computing, etc. The best is when the entire operation is a success and gives us the final routine; the worst is when the given method is clearly inadequate and we need to try another method. Sometimes, a particular sequencing of methods may be needed.

However, the preparation for computations can be different. Thus, in solving large-scale problems, the major problem is the system of filing the data, and this has the primary impact on the choice of the method as well as the program. When a microprocessor is used for optimization, the main issue is how simple and, also, how stable to errors the method is. The on-line processing of input data in time and using recurrently the resulting approximations are of vital importance. The man-machine mode opens several new options, viz. adjustment of algorithm parameters becomes simpler, the impact of the altered conditions on the procedure of solving the problem is possible to analyze, various methods can be tested sequentially, etc.

### 12.1.3 Evaluation of Solutions

The procedure of solving the problem does not terminate with the choice of the code and, next, the computations. It is useful to have an estimate of the accuracy of the solution obtained. In some cases this can be done in theoretical terms by using the duality theorem, checking the violation of the optimality condition, applying bounds similar to those in Theorem 2 of Section 1.5, etc. The accuracy can also be checked empirically, e.g., make a descent from differing initial points and see how different the results are. Furthermore, analyze the solution obtained, taking into consideration that the mathematical formulation of the problem was obtained after the primal problem had been made simpler, or coarser. It may happen as well that despite the fact that the point found is the solution of the mathematical model, it still does not satisfy the user, because some components have been missing in the initial formulation. (This is, in fact, the case with economic problems, and that is why optimization models are not common in economics modelling.) In this case the initial model must be reevaluated and modified, and the optimization problem be solved anew.

## 12.2 OPTIMIZATION PROGRAMS

### 12.2.1 General Requirements

Requirements to optimization programs in general depend, of course, on the purpose and application of these programs. Here are a few typical situations.

If the program is intended for a single program run, it can be arbitrary as long as it performs its function. For the case of multiple solutions of similar problems by only one customer, the computation time becomes an important factor. In particular, there is an appreciable payoff if the program is coded in machine language. Otherwise, however, similar programs for "internal use" do not need to obey particular rigid constraints.

Standard programs to be used by many customers in solving standard problems, must meet rigorous requirements, including:

- (a) be usable on computers of different modes (hence they are usually written in algorithmic languages, most frequently in FORTRAN);
- (b) be convenient in usage (this is of particular importance in linear programming problems with large data volume, in which a file building is not restricted);
- (c) be usable as subroutines (e.g., unconstrained minimization becomes frequently an auxiliary problem in solving the more general optimization problem);
- (d) provide a convenient and sufficiently detailed representation of the computational results.

The standard routine must contain descriptions and test examples. The customer does not need to be familiar with the minimization method in use and specify a great number of algorithm parameters. In general, the data provided by the user must be minimal (e.g., in unconstrained minimization problems, procedures for finding the function and the gradient, the initial point, and the computation accuracy test. When the solution cannot be found, the program must print out why. We have given only a partial list of requirements to standard optimization routines.

Today, widely used are optimization program packages. They are intended for solving large classes of problems and consist of related programs, with common modules (one-dimensional minimization type). They have the same requirements for data acquisition and print-out of the results. The program packages are more economical than a set of individual programs and more convenient for the user. They are a great asset for those who need to solve optimization problems routinely.

## 12.3 TEST PROBLEMS AND COMPUTATIONAL RESULTS

In this section we shall list several test optimization problems and the results obtained by means of some basic methods. They are useful in

evaluating the existing algorithms, as well as in developing new ones. We shall also discuss the criteria for a comparative analysis of the methods and different ways of representing the computational results.

### 12.3.1 Criteria for a Comparative Analysis of Algorithms. Empirical Results

It is not an easy task to compare several optimization methods on the basis of numerical experiments and rate one method over another. This arises from a number of factors:

- 1) it is not the methods that are compared but the computer realizations of the corresponding algorithms. A good method can be ruined, for example, by poor programming, by a wrong choice of the algorithm parameters, or by computing on a computer of smaller core memory;
- 2) there is no definite criterion as to how to evaluate the amount of “work” of each method. What seems natural to measure—the computing time—is not too convenient in practice. It is also hard to compare the speed of differing computers, or, it is sometimes impossible to know exactly the computing time (as in a multiprocessor mode), the computing time essentially depends on the programming language and the special features of the translator. A more reliable indicator is the the number of computations of the objective function or any other “intrinsic” characteristic of the method. A few problems arise, however, in this case. It is unclear how to compare the volume of computations for the function with that in solving various auxiliary problems (recall, for example, that the center-of-gravity method (9) in Section 5.4 is optimal with respect to the desired number of computations of the subgradient, but leads to a very complicated auxiliary problem in each iteration). It is also difficult to compare the computations of the function with those of its derivatives. In the problems where finite-difference approximations of the derivatives are applied, the answer is easier to obtain—one computation of the gradient is equivalent to  $n$  computations of the function, the computation of the Hessian is equivalent to  $n(n+1)/2$  computations of the function, etc. However, for a discrete optimal control problem the gradient is only about twice as costly as the function (see (33) of Section 11.2); for a quadratic function the gradient is even easier to compute than its values, the same applies to the computation of the subgradient in the minimax problem, etc.;
- 3) methods may behave differently at various stages of the minimization process. For example, let two methods be identical with respect to the volume of computations in each iteration, but for the first method the quantity  $\|x^k - x^*\|$  decreases rapidly at the beginning but hardly varies afterwards, whereas for the second method the same quantity varies slowly but at constant rate. Which of these two methods is better? Note also that quantities which characterize the accuracy of solution may vary as well (say,  $\|x^k - x^*\|$ ,  $\|\nabla f(x^k)\|$ ,  $f(x^k) - f^*$ , etc.).

There is no satisfactory way of resolving these problems. The only thing to do in this situation is to present the computational results in the detailed form, in order to compare the methods by the respective criteria.

The results of testing various methods cannot be meaningful unless the following rules are satisfied:

- 1) formulate the problem exactly; include all the parameters and initial approximation;
- 2) specify the type of computer, word length, programming language, compiler, program specifications;
- 3) describe in detail the algorithm, give the publication date when applicable;
- 4) derive intermediate results as well as the final one: say, every 100 iterations or for each increased order of accuracy;
- 5) describe the accuracy of each approximation ( $\|x^k - x^*\|$ ,  $f(x^k) - f^*$ , the residual in the constraints, and the extremum conditions satisfied); give the actual approximations  $x^k$  is problems of low dimensions;
- 6) indicate the volume of computations (the number of iterations, the number of calculations of the function  $f(x)$ ,  $\nabla f(x)$ , the computer time, etc.).

A similar detailed description of the methods is limited to publications devoted to a numerical testing, and not possible otherwise.

### 12.3.2 Test Problems: General Requirements

Frequently, authors suggest a new minimization method and construct only one or two examples to test it, without evaluating their algorithm versus the known ones. Or, they give computational results for some *a priori* inadequate methods. This seems to be typical of the publications on random search.

A comparative analysis of the methods requires the use of standard, preselected test problems, which satisfy the following requirements:

- 1) test problems must be standard and universal; they must be selected empirically, based on the fact how widely the particular method is used;
- 2) modelling of standard difficulties for a given class of problems (e.g., for unconstrained minimization problems, use tests with differing condition numbers, differing curvatures of level-lines, single modal and multimodal problems, etc.);
- 3) the solution to the test problem must be known;
- 4) the problems must be sufficiently brief (no large files or complicated rules for computing the functions);
- 5) do not use problems with specific features, advantaging some method versus the other; e.g., if we take  $f(x) = f_1(x_1) + \dots + f_n(x_n)$  as the test function for unconstrained minimization, the method of directional descent is extremely efficient; for a function of the form  $f(x) = F(\phi(x))$ , where  $\phi(x)$  is a quadratic function and  $F$  is a scalar function, the level sets are ellipsoids, and hence, say, Powell's method (see Lemma 2 in Section 3.4) are finite, and so on.

Standard, commonly used methods have not yet been developed. Nor is there a classification of test problems with respect to their complexity and potential difficulty. In what follows we shall make an attempt to list tests for major classes of optimization problems, which come maximally close to conditions 1-6.

### 12.3.3 Unconstrained Minimization of Smooth Functions

In the sequel we use the following notation:  $x = \{x_1, \dots, x_n\} \in \mathbf{R}^n$ ,  $f(x)$  is the objective function,  $x^*$  is a global minimum point of  $f(x)$  on  $\mathbf{R}^n$ ,  $f^* = f(x^*)$ ,  $x^0$  is the initial approximation,  $x^k$  is the resulting point,  $\delta f = f(x^k) - f^*$  is the solution accuracy in the function,  $\delta x = \max_{1 \leq i \leq n} |x_i^* - x_i^k|$  is the solution accuracy in the arguments,  $\mu$  is the condition number of  $x^*$  (see (4) of Section 1.3),  $k$  is the number of iterations of the method,  $k_0$  is the number of computations of the function,  $k_1$  is the number of computations of the gradient.

The simplest way to construct test problems with known solution is the following. Take a point  $x^*$  and functions  $\phi_i(x)$ ,  $i = 1, \dots, m$ . Then obviously

$$f(x) = \sum_{i=1}^m (\phi_i(x) - \phi_i(x^*))^2 \quad (1)$$

attains a minimum at  $x^*$ . Here  $f^* = 0$ ,  $\nabla^2 f(x^*) = \sum_{i=1}^m \nabla \phi_i(x^*) \nabla^T \phi_i(x^*)$ . Therefore we can change the condition number of the problem by appropriately choosing  $\phi_i(x)$ . In particular, if  $m < n$  or  $m = n$ , but  $\nabla \phi_i(x^*) = 0$  for some  $i$ , we obtain a singular minimum point. Functions of the form (1) are, in general, nonconvex and may have local (or even global) minima different from  $x^*$ .

Another way of constructing test problems with known solution involves choosing functions of the form

$$f(x) = \sum_{i=1}^{n+1} f_i(x_i, x_{i-1}) , \quad x_0 = a, \quad x_{n+1} = b . \quad (2)$$

Next, taking  $x_0^* = a$  and solving sequentially the one-dimensional equations

$$\partial f(x)/\partial x_i = \partial/\partial x_i [f_i(x_i^*, x_{i-1}^*) + f_{i+1}(x_{i+1}, x_i^*)] = 0$$

(denote the solution by  $x_{i+1}^*$ ) for  $i = 0, \dots, n$ , we take  $b = x_{n+1}^*$ . Clearly,  $\nabla f(x^*) = 0$ . If the  $f_i(x_i, x_{i-1})$  are convex in  $x_i, x_{i-1}$ , then  $f(x)$  is convex and  $x^*$  is a minimum point of  $f(x)$ .

Yet another technique involves taking an arbitrary smooth convex function  $\phi(x)$  and an arbitrary point  $x^*$  and construct

$$f(x) = \phi(x) - (\nabla \phi(x^*), x) . \quad (3)$$

Then  $\nabla f(x^*) = 0$ ,  $f(x)$  is convex and hence  $x^*$  is a global minimum point of  $f(x)$ . There are also many other ways of constructing functions with a known minimum point.

We shall use one of the foregoing techniques to construct the examples below. Note that special methods, with superior efficiency, exist for functions of the form (1), (2). For instance, for (1) the Gauss-Newton method (14) of Section 11.1 is efficient, for (2) the method of dynamic programming and the method of directional descent (36), (37) in Section 11.2. The specific features of problems (1), (2) may have an impact on the behavior of general minimization methods, as well.

We proceed to describe some particular test problems.

**Problem 1** (Rosenbrock's function (problem 2 in [0.21])):  $n = 2$ :

$$f(x) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2,$$

$$x^0 = (-1.2; 1), \quad x^* = (1; 1), \quad f^* = 0.$$

The function is ill-conditioned ( $\mu = 2,500$ ), nonconvex, with parabolic gully, the point  $x^0$  being far from  $x^*$  (Fig. 44). A possible multidimensional extension of problem 1 can be found in [12.3]:

$$f(x) = 100 \sum_{i=2}^n (x_i - x_{i-1}^2)^2 + (1 - x_1)^2,$$

$$x^* = (1; \dots; 1), \quad f^* = 0, \quad x_1^0 = -1, \quad x_i^0 = (x_{i-1}^0)^2 - 0.2,$$

$$i = 2, \dots, n.$$

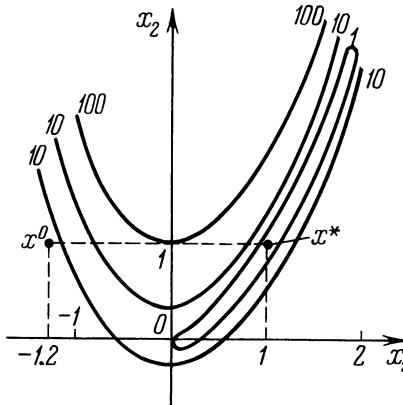


Fig. 44 Rosenbrock's function.

**Problem 2** (Powell's function (problem 26 in [0.21])):  $n = 4$ :

$$f(x) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4,$$

$$x^0 = (3; -1; 0; 1), \quad x^* = (0; 0; 0; 0), \quad f^* = 0.$$

The function is nonconvex, the minimum point is singular ( $\mu = \infty$ ).

**Problem 3** (a difference analog of the brachistochrone problem [3.9]):  $n > 1$  is arbitrary:

$$f(x) = \sum_{i=1}^{n+1} \left[ \frac{0.0016 + (x_i - x_{i-1})^2}{0.04 i} \right]^{1/2}, \quad x_0 = x_0^* = 0, \quad x_{n+1} = x_{n+1}^*,$$

$$x^0 = (0, \dots, 0), \quad x_{i+1}^* = x_i^* + 0.04 \left[ \frac{0.0099099(i+1)}{1 - 0.0099099(i+1)} \right]^{1/2}, \quad i = 0, \dots, n-1$$

This function is convex (even strongly convex on any bounded set, although not on the entire space), the condition number grows with  $n$ , but not too fast ( $\mu \sim 10^4$  for  $n = 50$ ).

**Problem 4** (mean square approximation by exponentials [3.9]):  $n = 4$ :

$$f(x) = \alpha^{-2} \sum_{j=1}^{10} [\alpha \exp(-0.2j) + 2\alpha \exp(-0.4j) - x_1 \exp(-0.2jx_2) - x_3 \exp(-0.2jx_4)]^2, \quad f^* = 0,$$

$$(a) \quad x^0 = (0.5; 0; 2.5; 3), \quad x^* = (1; 1; 2; 2), \quad \alpha = 1,$$

$$(b) \quad x^0 = (500; 0; 2500; 3), \quad x^* = (1000; 1; 2000; 2), \quad \alpha = 1000.$$

The function  $f(x)$  is nonconvex, has a twisted gully, the condition number is great, especially for 4(b).

**Problem 5** (mean square approximation by polynomials [11.22]):  $n$  is arbitrary:

$$f(x) = \sum_{j=1}^{101} \left( \sum_{i=1}^n x_i t_j^{i-1} - \sum_{i=1}^n x_i^* t_j^{i-1} \right)^2,$$

$$t_j = 1.01(j-1), \quad j = 1, \dots, 101, \quad x^0 = (2; \dots; 2), \quad x^* = (1; \dots; 1), \quad f^* = 0.$$

The function  $f(x)$  is quadratic; its condition number is extremely large (in Subsection 11.1.5, the condition numbers were given for the continuous variant of the problem:  $\mu = 1.5 \cdot 10^7$  for  $n = 6$ ,  $\mu = 1.6 \cdot 10^{13}$  for  $n = 10$ ; we may expect that the condition number for a discrete problem is of the same order).

TABLE 1  
PRIMAL METHODS FOR PROBLEMS 1-3

METHOD	Problem 1		
	$k_0$	$\delta x$	$\delta f$
<i>Gradient</i>	7657	$0.2 \cdot 10^{-5}$	$10^{-9}$
<i>Barycenter</i>	674	$10^{-9}$	$10^{-17}$
<i>Powell</i>	151	—	$10^{-10}$
<i>Simplex</i>	200	—	$10^{-8}$

METHOD	Problem 2		
	$k_0$	$\delta x$	$\delta f$
<i>Gradient</i>	N	N	N
<i>Barycenter</i>	924	$10^{-4}$	$10^{-18}$
<i>Powell</i>	433	—	$10^{-13}$
<i>Simplex</i>	209	—	$0.7 \cdot 10^{-7}$

Problem 3: $n = 50$			
METHOD	$k_0$	$\delta x$	$\delta f$
<i>Gradient</i>	N	N	N
<i>Barycenter</i>	21,721	$10^{-4}$	$10^{-7}$
<i>Powell</i>	—	—	—
<i>Simplex</i>	—	—	—

V.A. Skokov and Yu.E. Nesterov numerically tested the methods, using test problems 1-5; here is a review of their results.

Table 1 lists data for problems 1-3 and four primal methods (i.e., no computation of the derivatives, see Section 3.4): 1) a difference analog of the gradient method (9) in Section 3.4 (*Gradient*); 2) a combination of the methods of directional descent (11) of Section 3.4, a difference variant of the conjugate gradient method (24) of Section 3.2 and the method of barycentric coordinates (19) of Section 3.4 (*Barycenter*, see [12.2]); 3) Powell's method of Subsection 3.4.4 (*Powell*); and 4) the simplex method (18) of Section 3.4 (*Simplex*). N means that the method has led to no solution, dash (–) means no computations have been made.

As is seen from Table 1, the gradient method does not operate even for problems of small ( $n = 14$ ) dimension. The remaining methods are, roughly, equivalent.

Table 2 lists the computational results (for the same problems 1-3) using first-order methods (i.e., with  $\nabla f(x)$ ). Computations were made for the following methods: 1) the steepest descent variant of the gradient method (3), (4) in Section 3.1 (*Gradient*); 2) two versions of the conjugate gradient method (22) and (24) of Section 3.2 (*Congrad-1* and *Congrad-2*, respectively); 3) Davidon-Fletcher-Powell method (8) of Section 3.3 (*DFP*); 4) Broyden-Fletcher-Shanno method (10) of Section 3.3 (*BFS*); and 5) Shor's method (14), (17) of Section 5.4 (*Shor-1*) based on the data of [5.15]. The results in Table 2 show once again that the gradient method is not usable for solving even relatively simple problems. The quasi-Newton methods (especially *BFS*) perform somewhat better than the conjugate gradient method; however, they require a larger computer memory and a large number of operations in each iteration. The *Congrad-2* version is marginally more efficient. Shor's method is of superior efficiency; it is as good as the best quasi-Newton methods. Since Shor's method is equally usable for the minimization of non-smooth functions (for which it was originally developed, see Section 5.4), it can be regarded to be efficient as well as universal.

Problem 3 was investigated numerically in [Section 31 in 0.19], with the aim of obtaining the most exact solution of a continuous brachistochrone problem. In this respect, Problem 3 is a poor discrete approximation of the continuous problem; nevertheless, it is a test problem of finite-dimensional minimization on its own merit. In [0.19] the author shows how to construct a rapidly converging minimization method which uses the special features of problem 3. One needs to bear in mind that special highly efficient methods can be found for each concrete test problem. The purpose of using test problems is, however, to have a tool for testing standard minimization methods.

For problem 4 the following methods were compared: 1) the method *Congrad-2*; 2) the method *Congrad-3*, that is the conjugate gradient method combined with the variable metric method which was suggested in [3.9]; 3) the method *Shor-1*; 4) a *Gauss-Newton* method (14) of Section 11.1, as

described in [11.22]. The results are listed in Table 3, where we can see that for problem 4(a) the conjugate gradient method converges slowly, and for the more difficult problem 4(b) it does not lead to a solution at all. The modification *Congrad-3* of this method is much more efficient. Shor's method is quite efficient for these problems. A special Gauss-Newton method works much better than the general methods, which is attributed to a great extent to the special features of the problem, viz. the efficiency of the Gauss-Newton method increases as  $f^*$  decreases, see Section 11.1; in the cases cited in Table 3 we have  $f^* = 0$ . Note also that for this method the volume of computations per each iteration is essentially larger than that for other methods.

TABLE 2  
FIRST-ORDER METHODS FOR PROBLEMS 1-3

METHOD	Problem 1				Problem 2			
	$k$	$k_1$	$\delta f$	$\delta x$	$k$	$k_1$	$\delta f$	$\delta x$
Gradient	217	476	$10^{-4}$	$10^{-2}$	1000	2003	$10^{-3}$	$10^{-1}$
Congrad-1	32	65	$10^{-9}$	$10^{-5}$	39	82	$10^{-5}$	$10^{-2}$
Congrad-2	30	63	$10^{-9}$	$10^{-5}$	27	59	$10^{-4}$	$10^{-2}$
DFP	20	87	$10^{-9}$	$10^{-5}$	16	54	$10^{-9}$	$10^{-3}$
BFS	22	72	$10^{-11}$	$10^{-6}$	16	54	$10^{-9}$	$10^{-3}$
Shor-1	39	$k_0=286$	$10^{-15}$	$10^{-7}$	50	$k_0=695$	$10^{-13}$	$10^{-6}$

TABLE 3  
METHODS FOR PROBLEM 4

METHOD	Problem 4(a)				Problem 4(b)			
	$k=k_1$	$k_0$	$\delta f$	$\delta x$	$k=k_1$	$k_0$	$\delta f$	$\delta x$
<i>Congrad-2</i>	179	781	$10^{-13}$	$10^{-4}$	N	N	N	N
<i>Congrad-3</i>	47	205	$10^{-18}$	$10^{-8}$	110	622	$10^{-18}$	$4 \cdot 10^{-4}$
<i>Shor-1</i>	83	400	$6 \cdot 10^{-5}$	$10^{-7}$	72	410	$4 \cdot 10^{-14}$	$10^{-3}$
<i>Gauss-Newton</i>	6	—	$10^{-21}$	$10^{-11}$	11	—	$10^{-22}$	$10^{-7}$

TABLE 4  
METHODS FOR PROBLEM 5

METHOD	$n$	2	3	4	5
		$\delta f$	0	0	0
<i>Gauss-Newton</i>		$\delta f$	0	0	$2 \cdot 10^{-17}$
<i>LSM</i>		$\delta f$	0	0	$3 \cdot 10^{-18}$
<i>Gauss-Newton</i>		$\delta x$	$10^{-8}$	$10^{-8}$	$10^{-8}$
<i>LSM</i>		$\delta x$	$10^{-8}$	$10^{-8}$	$6 \cdot 10^{-8}$
					$2 \cdot 10^{-6}$
METHOD	$n$	6	7	8	9
		0	0	0	0
<i>Gauss-Newton</i>		$\delta f$	0	0	0
<i>LSM</i>		$\delta f$	$8 \cdot 10^{-17}$	$2 \cdot 10^{-16}$	$6 \cdot 10^{-16}$
<i>Gauss-Newton</i>		$\delta x$	$10^{-8}$	$10^{-7}$	$10^{-6}$
<i>LSM</i>		$\delta x$	$10^{-4}$	$3 \cdot 10^{-3}$	0.26
					3.8
					5.1

For problem 5 the following two methods were compared: the variant of the Gauss-Newton method of [11.22] cited above and the usual least squares method (*LSM*). The system of linear equations  $\nabla f(x) = 0$  was solved by the Gauss elimination method; the results are listed in Table 4. For  $n \leq 10$  both methods yield an almost exact minimum in the function; the accuracy in the argument is significantly higher for the iterative method. However, the computer time for the Gauss-Newton method is roughly 10 times greater than for the least squares method.

### Exercises

1. What is the way to construct a test problem of the form  $f(x) = \sum_{i=1}^m ((a^i, x) - b_i)^2$ , in which  $x^*$  is known but  $f^* > 0$ ?
2. Do you think the quantities  $\delta x$  in Table 4 are large or small with respect to the theoretical values for the given condition number  $\mu$ ?

#### 12.3.4 Unconstrained Minimization of Nonsmooth Functions

Nonsmooth test functions of the form

$$f(x) = \max_{1 \leq i \leq m} f_i(x), \quad (4)$$

with known minimum, can be constructed in the following way. Take convex smooth functions  $\phi_i(x)$ ,  $i = 1, \dots, m$  (say, quadratic or linear functions) and a point  $x^* \in \mathbf{R}^n$ . Calculate the gradients  $\nabla \phi_i(x^*)$  and find the set of indices  $I$  and numbers  $\lambda_i > 0$ ,  $i \in I$ , such that  $\sum_{i \in I} \lambda_i \nabla \phi_i(x^*) = 0$  (this can be done by adding to the set  $\phi_i(x)$  linear functions of the form  $(x, e_j)$ , where the  $e_j$  are standard basis vectors). Next, take  $f_i(x) = \phi_i(x) - \alpha_i$ , where  $\alpha_i = \phi_i(x^*)$  for  $i \in I$  and  $\alpha_i > \phi_i(x^*)$  for  $i \notin I$ ; then, according to the extremum conditions (see Lemma 11 of Section 5.1), the function  $f(x)$  of the form (4) has a global minimum at  $x^*$ . There are other procedures, too, for constructing test functions (see, e.g., problem 8 below).

**Problem 6** (Shor's function [5.15, p. 176]):  $n = 5$ :

$$f(x) = \max_{1 \leq i \leq 10} b_i \|x - a^i\|^2, \quad b = (1; 5; 10; 2; 4; 3; 1.7; 2.5; 6; 3.5)$$

$$A = \begin{pmatrix} 0 & 2 & 1 & 1 & 3 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 2 & 4 & 2 & 2 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & \\ 0 & 1 & 1 & 2 & 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 3 & 2 & 2 & 1 & 1 & 1 & 0 & 0 & \end{pmatrix},$$

$$x^* = (1.12434; 0.97945; 1.47770; 0.92023; 1.12429),$$

$$f^* = 22.60016, \quad I = (2, 4, 5, 9), \quad x^0 = (0; 0; 0; 0; 1).$$

Here  $A$  is the matrix with columns  $a^i$ ,  $I$  is the set of active indices:  $I = \{i : b_i \|x^* - a^i\|^2 = f(x^*)\}$ .

In [5.15], a modification of the *Shor-1* method is used (i.e., method (14), (17) of Section 5.4), with a special technique of choosing the step size  $\gamma_k$ ; 51 iterations (51 calculations of  $\delta f(x)$ ) yielded  $\delta f = 7 \cdot 10^{-4}$ ,  $\delta x = 4 \cdot 10^{-4}$ . For the same problem the author of [0.19, pp. 415-7] employed a variant of the linearization method; it took 14 times to solve the auxiliary linear programming problem in order to reach the accuracy  $\delta f = 5 \cdot 10^{-5}$ .

**Problem 7** [5.20, p. 151]:  $n = 10$ :

$$f(x) = \max_{1 \leq k \leq 5} \{(A^{(k)}x, x) - (b^k, x)\},$$

$$A_{ij}^{(k)} = \exp(i/j) \cos ij \sin k, \quad i \neq j,$$

$$A_{ii}^{(k)} = 0.1i \sin k + \sum_{j \neq i} A_{ij}^{(k)},$$

$$b_i^k = \exp(i/k) \sin ik, \quad i, j = 1, \dots, 10, \quad k = 1, \dots, 5,$$

$$x^* = (-0.1263; -0.0346; -0.0067; 0.2668; 0.0673; 0.2786; 0.0744; \\ 0.1387; 0.0839; 0.0385),$$

$$f^* = -0.8414, \quad x^0 = (1; \dots; 1), \quad f^0 = 5337.$$

C. Lemarechal obtained the following. The *Shor-1* method gave  $\delta f < 10^{-4}$  in  $k_1 = 60$  (i.e., for 60 calculations of the subgradient). Two other nonsmooth minimization methods of P. Wolfe and C. Lemarechal [5.20] required  $k_1 = 186$  and  $k_1 = 96$  to obtain the same accuracy. The attempt to apply one of the best methods among quasi-Newton methods of smooth minimization, the *BFS* method, did not give an accurate solution: in  $k_1 = 86$  the result was  $\delta f = 0.08$ , after which the method was stopped.

**Problem 8** (approximation by polynomials in  $\ell_1$ ),  $n$  is arbitrary:

$$f(x) = \sum_{j=1}^{101} \left| \sum_{i=1}^n x_i t_j^{i-1} - \sum_{i=1}^n x_i^* t_j^{i-1} \right|,$$

$$t_j = 0.01(j-1), \quad j = 1, \dots, 101,$$

$$x^* = (1/n; \dots; 1/n), \quad f^* = 0, \quad x^0 = (0; \dots; 0).$$

 For  $n = 20$  a point with  $\delta f = 10^{-7}$  was found after 180 iterations; however

This problem is a nonsmooth analog of problem 5. In a number of methods it is required to specify the region  $Q_0$  of localization of the minimum; in the given case  $Q_0 = \{x: \|x\| \leq 1\}$  was chosen. V.I. Venetz, O.V. Gulinskij, and A.S. Nemirovskij carried out computations for differing algorithms. The ellipsoid method (12) in Section 5.4 converged (in the function) for large-scale problems, up to and including  $n = 50$ . It is worth noting that this method attained a relatively high accuracy level. For example, for  $n = 5$  in 1,200 iterations the results are  $\delta f = 5 \cdot 10^{-11}$ ,  $\delta x = 10^{-9}$ ; for  $n = 20$  in 9,500 iterations the results are  $\delta f = 10^{-11}$ ,  $\delta x = 10^{-3}$ . The *Shor-1* method ((14), (17) of Section 5.4) converged significantly faster; for example, for  $n = 5$  only 47 iterations were needed to obtain  $\delta f = 4 \cdot 10^{-4}$ ,  $\delta x = 10^{-5}$  (of course in the latter case the accuracy in the argument was very low ( $\delta x = 1$ )). The rate of convergence depended essentially on  $\alpha$  (see (17), (19) of Section 5.4); since the function  $f(x)$  is homogeneous with exponent 1, one can take  $M = N = 1$ ,  $\alpha = \infty$  and (theoretically) make the method finite. Venetz, Gulinskij, and Nemirovskij tested method (14), (18) of Section 5.4 (*Shor-2*). This method is easier to apply to the present problem since the value  $f^* = 0$  is known. If  $f^*$  is assumed unknown, regulation of the step size in the *Shor-2* method becomes complicated, and the rate of convergence may decrease considerably.

### 12.3.5 Nonlinear Programming

We shall briefly describe the way of constructing convex programming test problems, with inequality constraints. Take a point  $x^* \in \mathbf{R}^n$ , a set of convex differentiable functions  $g_i(x)$ ,  $i = 0, \dots, m$ , and a set of indices  $I^*$  of cardinality  $\leq n$ . Next we construct the problem

$$\begin{aligned} & \min [g_0(x) + (c - g_0(x^*), x)] , \\ & g_i(x) - g_i(x^*) \leq 0 , \quad i \in I^* , \\ & g_i(x) - \alpha_i \leq 0 , \quad i \in \bar{I^*} , \end{aligned} \tag{5}$$

where  $\alpha_i > g_i(x^*)$  are arbitrary numbers and the vector  $c$  has the form

$$c = - \sum_{i \in I^*} \lambda_i \nabla g_i(x^*) , \tag{6}$$

$\lambda_i \geq 0$ ,  $i \in I^*$ , are arbitrary numbers. Then, according to the Kuhn-Tucker theorem,  $x^*$  is a solution of problem (5). By varying the number of elements in  $I^*$  one can vary the dimension of the manifold on which the solution is obtained. In particular, if  $I^*$  contains  $n$  elements so that the gradients  $\nabla g_i(x^*)$ ,  $i \in I^*$ , are linearly independent,  $\lambda_i > 0$ ,  $i \in I^*$ , then  $x^*$  is a sharp minimum point (see Theorem 4 of Section 9.2).

Let us introduce some additional notation:

$m$  is the total number of constraints (excluding those of the form  $x \geq 0$ );  
 $r$  is the number of inequality constraints;

$\delta g$  is the residual of the resulting solution in the constraints:

$$\delta g = \sum_{i=1}^r g_i(x^k)_+^2 + \sum_{i=r+1}^m g_i(x^k)^2,$$

$k_0$  and  $k_1$  are respectively the number of calculations of all the functions and of the gradients; e.g., the number of calculations of the augmented Lagrangian and of its gradient.

**Problem 9** (problem 18 in [0.21]):  $n = 15$ ,  $m = 5$ :

$$\begin{aligned} f(x) &= \sqrt{\sum_{i=1}^{10} b_i x_i + \sum_{i,j=1}^5 c_{ij} x_{10+i} x_{10+j} + 2 \sum_{i=1}^5 d_i x_{10+i}^3}, \\ g_i(x) &= \sum_{j=1}^{10} a_{ij} x_j - 2 \sum_{j=1}^5 c_{ij} x_{10+j} - 3d_i x_{10+i}^2 - e_i \leq 0, \quad i = 1, \dots, 5, \end{aligned}$$

$$x_i \geq 0, \quad i = 1, \dots, 15,$$

$$C = \begin{pmatrix} 30 & -20 & -10 & 32 & -10 \\ -20 & 39 & -6 & -31 & 32 \\ -10 & -6 & 10 & -6 & -10 \\ 32 & -31 & -6 & 39 & -20 \\ -10 & 32 & -10 & -20 & 30 \end{pmatrix},$$

$$A = \begin{pmatrix} -16 & 0 & -3.5 & 0 & 0 & 2 & -1 & -1 & 1 & 1 \\ 2 & -2 & 0 & -2 & -9 & 0 & -1 & -2 & 2 & 1 \\ 0 & 0 & 2 & 0 & -2 & -4 & -1 & -3 & 3 & 1 \\ 1 & 0.4 & 0 & -4 & 1 & 0 & -1 & -2 & 4 & 1 \\ 0 & 2 & 0 & -1 & -2.8 & 0 & -1 & -1 & 5 & 1 \end{pmatrix},$$

$$b = (-40; -2; -0.25; -4; -4; -1; -40; -60; 5; 1),$$

$$d = (4; 8; 10; 6; 2),$$

$$e = (-15; -27; -36; -18; -12),$$

$$x_i^0 = 0.0001, \quad i = 1, \dots, 15, \quad i \neq 7, \quad x_7^0 = 60, \quad f^0 = 2400.01.$$

In this problem both the objective function and the constraints are nonlinear and nonconvex. The exact solution is unknown. In [0.21] the following approximate solution is given:  $x^* = (0; 0; 5.174; 0; 3.0611; 11.8395; 0; 0; 0.1039; 0; 0.3; 0.3335; 0.4; 0.4283; 0.224)$ ,  $f^* = 32.386$ . However,

in [12.5] a somewhat different solution is obtained with a smaller value of the function:  $x^* = (0; 0; 5.174; 0; 3.061117; 11.839466; 0; 0; 0.103877; 0; 0.300002; 0.333466; 0.400003; 0.428306; 0.223964), f^* = 32.348679.$

Problem 9 is a problem of medium difficulty. It has been used by some authors (see bibliography in [0.21] and [12.5]) to compare different methods. According to the results obtained in [0.21, Tables 9.3.5 and 9.3.6], no solution of the problem was obtained when the following methods were used: the sliding tolerance method; the linearization method (program POP); Rosenbrock's method; and the generalized gradient method (GGMOP) (for the terminology see [0.21]). Variants of the barrier-function method as well as of penalty-function method for  $k_1$  of the order of several thousands yield only a sufficiently rough solution  $\delta x \sim 0.1/10.3$  and  $\delta f \sim 10^{-2}/10^{-3}$ . Table 5 lists I ...  
the computation results obtained in [12.5] for the penalty-function method (10) of Section 9.4 and the method of the augmented Lagrangian (23) of Section 9.3 for three initial points  $x^0$ :

- (a)  $x_i^0 = 0, i = 1, \dots, 10, i \neq 7; x_7^0 = 60; x_i^0 = 0.0001, i = 11, \dots, 15; x_i^0 = 0, i = 16, \dots, 20;$
- (b)  $x_i^0 = 1, i = 1, \dots, 15; x_i^0 = 0, i = 16, \dots, 20;$
- (c)  $x_i^0 = 1, i = 1, \dots, 15; x_i^0 = 10, i = 16, \dots, 20.$

Here the vector  $x$  has 20 coordinates since the problem is written in the form

$$\min f(x),$$

$$g_i(x) + x_{15+i} = 0, \quad i = 1, \dots, 5, \quad x_i \geq 0, \quad i = 1, \dots, 20,$$

that is the canonical form for the programs used in [12.5].

The data in Table 5 show that descent from essentially distinct initial points is approximately identical, and the method of the augmented Lagrangian is appreciably more efficient than the penalty-function method. Furthermore,  
<sup>2p</sup>

TABLE 5  
METHODS FOR PROBLEM 5

Initial point	$\delta x$	$\delta f$	$\delta g$	$k_0$	$k_1$
The method of the augmented Lagrangian					
(a)	$10^{-6}$	$10^{-6}$	$0.5 \cdot 10^{-12}$	1199	3921
(b)	$10^{-4}$	$10^{-6}$	$0.4 \cdot 10^{-11}$	1248	4256
(c)	$10^{-5}$	$10^{-6}$	$0.5 \cdot 10^{-10}$	1488	4911

(Table 5 continued)

## The penalty-function method

(a)	$10^{-2}$	$0.4 \cdot 10^{-3}$	$0.2 \cdot 10^{-6}$	1776	6151
(b)	$10^{-2}$	$0.6 \cdot 10^{-3}$	$0.3 \cdot 10^{-6}$	1188	4135
(c)	$2 \cdot 10^{-2}$	$0.3 \cdot 10^{-3}$	$0.2 \cdot 10^{-6}$	1078	3895

---

the penalty coefficient  $K$  is regulated and its final value becomes equal to  $\sim 100$  in the former method and to  $\sim 1,000$  in the latter method, respectively. The computation for one particular variant took about three minutes on a BESM-6 computer in the BESM-ALGOL.

**Problem 10** ([0.21, problem 20]):  $n = 24$ ,  $r = 14$ ,  $m = 20$ :

$$\begin{aligned}
 f(x) &= \sum_{i=1}^{12} a_i(x_i + x_{i+12}) , \\
 g_i(x) &= \frac{x_{i+12}}{b_i \sum_{j=13}^{24} x_j} - \frac{c_i x_i}{40 b_i \sum_{j=1}^{12} \frac{x_j}{b_j}} = 0 , \quad i = 1, \dots, 12 , \\
 g_{13}(x) &= \sum_{i=1}^{24} x_i - 1 = 0 , \\
 g_{14}(x) &= \sum_{i=1}^{12} \frac{x_i}{d_i} + 142.224705 \sum_{i=13}^{24} \frac{x_i}{b_{i-12}} - 1.671 = 0 , \\
 g_{15}(x) &= \frac{x_1 + x_{13}}{\sum_{i=1}^{24} x_i} - 0.1 \leq 0 , \\
 g_{16}(x) &= \frac{x_2 + x_{14}}{\sum_{i=1}^{24} x_i} - 0.3 \leq 0 , \\
 g_{17}(x) &= \frac{x_3 + x_{15}}{\sum_{i=1}^{24} x_i} - 0.4 \leq 0 , \\
 g_{18}(x) &= \frac{x_7 + x_{19}}{\sum_{i=1}^{24} x_i} - 0.3 \leq 0 , \\
 g_{19}(x) &= \frac{x_8 + x_{20}}{\sum_{i=1}^{24} x_i} - 0.6 \leq 0 ,
 \end{aligned}$$

$$g_{20}(x) = \frac{x_9 + x_{21}}{\sum_{i=1}^{24} x_i} - 0.3 \leq 0,$$

$$x_i \geq 0, \quad i = 1, \dots, 24,$$

$$a = (0.0698; 0.0577; 0.05; 0.2; 0.26; 0.55; 0.06; 0.1; 0.12; 0.18; 0.1; 0.09), \quad \text{F3}$$

$$b = (44.94; 58.12; 58.12; 137.4; 120.9; 170.9; 62.501; 84.94; 133.425; 72.507; 46.07; 60.097), \quad \text{F9}$$

$$c = (123.7; 31.7; 45.7; 14.7; 84.7; 27.7; 19.7; 7.1; 2.1; 17.7; 0.85; 0.64), \quad \text{F8}$$

$$d = (31.244; 36.12; 34.784; 92.7; 82.7; 91.6; 56.708; 82.7; 80.8; 64.517; 49.4; 49.1) \quad \text{F4}$$

$$x_i^{\sqrt{}} = 0.04, \quad i = 1, \dots, 24.$$

V(0)

The problem has nonconvex and nonlinear constraints. It is possible to slightly simplify the problem by substituting 1 for  $\sum_{i=1}^{24} x_i$  in the constraints 15 through 20; this is not permitted, however, in testing general methods. The solutions obtained in [0.21] are coarse; the most accurate solution has been obtained in [12.5]:  $x^* = (0; 0.107248; 0.111390; 0; 0; 0; 0.0755407; 0; 0; 0; 0.0111949; 0; 0.192752; 0.288611; 0; 0; 0; 0.212858; 0; 0; 0; 0)$ ,  $f^* = 0.0556580$ .

Among the problems listed in [0.21], this problem is the most complex; among standard problems, none led to a solution. According to the data therein, only the reduced gradient method and the sliding tolerance method proved efficient; the latter method required 8 min 30 sec of computations on a CDC-6600 to obtain the accuracy of  $\delta f \sim 10^{-3}$ ,  $\delta x \sim 0.4 \cdot 10^{-1}$ . In [12.5] the problem was solved by the augmented Lagrangian method and by the penalty-function method, with respective results:  $\delta x = 0.5 \cdot 10^{-6}$ ,  $\delta f = 10^{-8}$ ,  $\delta g = 10^{-13}$  for  $k_0 = 1,828$ ,  $k_1 = 5,436$  and 5 min on a BESM-6, and  $\delta x = 0.3 \cdot 10^{-3}$ ,  $\delta f = 0.3 \cdot 10^{-4}$ ,  $\delta g = 0.3 \cdot 10^{-4}$ ,  $k_0 = 204$ ,  $k_1 = 600$ . A higher accuracy could be achieved by continuing the penalty-function method. To conclude, the penalty-function method is appropriate for a coarse solution, the augmented Lagrangian to be applied for subsequent refinement.

### Exercises

3. Construct the dual to problem 9. Show that it can be written in the form

$$\min \left[ (e, y) + (Cy, y) + \sum_{i=1}^5 d_i y_i^3 \right], \quad y \in \mathbb{R}^5, \quad A^T y \geq b, \quad y \geq 0,$$

is convex, and its solution is  $y^* = (0.3; 0.3335; 0.4; 0.4285; 0.224)$ .

4. What is the way to construct a test problem similar to (5), in which all the constraints are equalities?

### 12.3.6 Linear Programming

However strange it may seem, there are no commonly accepted linear programming problems of graduated dimension and complexity. As a rule, new routines are first tested on “toy” problems of small dimension (to see that they work properly), and next applied to actual problems. At this stage no attempt is usually made to compare the efficiency of programs on the basis of identical numerical material. It only makes sense to make such a comparison using problems where  $m \geq 50$ ,  $n \geq 100$  (in the canonical notation). Problems of smaller dimension can, as a rule, be solved by good programs of the simplex method on most recent computers in a negligibly short computer time (of the order of seconds). The coding in problems of this dimension is, however, quite labor consuming, and there is no reason to waste time unless absolutely essential.

Apparently, it is advantageous to have linear programming test problems of two types: special difficult small-scale problems and large-scale problems in which the constraint matrix is easily generated. The class of the former problems includes (i) the examples constructed, in which the simplex method requires a great number of steps (depending exponentially on the dimension; see problem 11 below) and (ii) unstable problems in which the matrices of the linear equations to be solved at each step of the simplex method are ill-conditioned. The latter can be obtained by using discrete problems of uniform  $\ell_1$ -norm approximation of a function by polynomials written in the form of linear programming (see (19) of Section 11.1).

The latter class of test problems can be constructed, for example, in the following way. Take any simple convex nonlinear function  $f(x)$  in  $\mathbb{R}^n$  and collection of points  $x_1, \dots, x^m$ , including the minimum point  $x^*$  of  $f(x)$ . Then the linear programming problem

$$\begin{aligned} & \min t, \\ & (\nabla f(x^i), x - x^i) - t \leq -f(x^i), \quad i = 1, \dots, m, \end{aligned} \tag{7}$$

has a solution  $\{x^*, 0\}$  (cf. the methods for minimizing  $f(x)$  described in Section 5.4). Thus, if we take  $f(x) = \|x\|^2$  and for  $x^i$  points of the form  $\{x_1, \dots, x_n\}$ , where the  $x_j$ ,  $j = 1, \dots, n$ , take on the values 0,  $\pm 1$ , we obtain the linear programming problem

$$\begin{aligned} & \min t, \\ & 2(x^i, x) - t \leq \|x^i\|^2, \end{aligned} \tag{8}$$

with  $3^n$  constraints and the solution is  $x^* = 0$ ,  $t^* = 0$ . As far as we know,

no one has made a numerical study of similar multidimensional problems which can be so easily formulated (constraint matrix easily generated).

**Problem 11:**  $n$  is arbitrary:

$$\begin{aligned} \max x_n \\ -1 &\leq x_1 \leq 1, \\ -2x_{i-1} - 2^{2^{i-2}} &\leq x_i \leq 2x_{i-1} + 2^{2^{i-2}}, \quad i = 2, \dots, n, \\ x_1^* &= 1, \quad x_i^* = 2x_{i-1}^* + 2^{2^{i-2}}, \quad i = 2, \dots, n. \end{aligned} \tag{9}$$

In this problem the constraint polyhedron has  $2^n$  vertices; if we start from the vertex

$$\begin{aligned} x_1^0 &= 1, \quad x_i^0 = 2x_{i-1}^0 + 2^{2^{i-2}}, \quad i = 2, \dots, n-1, \\ x_n^0 &= -2x_{n-1}^0 - 2^{2^{n-2}}, \end{aligned}$$

then the iterations of the simplex method for one of the procedures for passing to a neighboring vertex will run through all the vertices of the polyhedron (Exercise 6).

We can pose the problem in the canonical form as follows:

$$\begin{aligned} \max & \sum_{j=1}^m c_j x_j, \\ \sum_{j=1}^i a_{ij} x_j + x_{n+i} &= b_i, \quad i = 1, \dots, n, \\ x_j &\geq 0, \quad j = 1, \dots, 2n, \\ a_{ii} &= 1, \quad a_{ij} = (-1)^{i+j} 2^{i-j+1}, \quad j < i, \\ b_i &= \frac{2^{2i} - (-2)^i}{3}, \quad i = 1, \dots, n, \\ c_j &= (-2)^{n-j}, \quad j = 1, \dots, n. \\ x_i^0 &= 0, \quad x_{n+i}^0 = b_i, \quad i = 1, \dots, n. \end{aligned} \tag{10}$$

Here  $x_i^*$ ,  $i = 1, \dots, n$ , are the same as before;  $x_i^* = 0$ ,  $i = n+1, \dots, 2n$ ; the solution of the dual problem has the form  $y_i^* = 2^{n-i}$ ,  $i = 1, \dots, n$ . A numerical study of this problem was made by N.A. Sokolov. In the practical implementation of the simplex method, for  $n = 15$  the number of steps was 21,402 (instead of the theoretical number  $2^{15} = 32,768$ ); apparently, this is due

to the method of choosing the adjacent vertex in this particular algorithm. Even so, this is a very *large* number of steps (for problems of similar dimension the number of iterations does not usually exceed 100). Solution by means of an iteration method based on the augmented Lagrangian (like (12) in Section 10.3) yielded accuracy of roughly 2 to 3 percent (both in the arguments and in the criterion) in 556 iterations, where each iteration included roughly 7 cycles of the directional descent method for minimizing the augmented Lagrangian.

### Exercises

5. Show that the following problem [10.18] is equivalent to (9):

$$\begin{aligned} & \max x_n , \\ & 0 \leq x_1 \leq 1 , \\ & \varepsilon x_{i-1} \leq x_i \leq 1 - \varepsilon x_{i-1} , \quad i = 2, \dots, n , \quad 0 < \varepsilon < \frac{1}{2} , \\ & x^0 = (0; \dots; 0) , \quad x^* = (0; \dots; 0; 1) . \end{aligned}$$

6. Find the path through  $2^n$  adjacent vertices of the polyhedron (9) for which  $x^n$  increases monotonically.

## NOTES

### Chapter 1

**1.1.** The basic facts concerning a differentiation of functions of several variables can be found in any standard Analysis text, e.g. Shilov [1.10] or Goldstein [1.14]. Differentiation of functionals and operators are treated in Kantorovich and Akilov [1.5] and Vajnberg [1.2].

Properties of convex and strongly convex functions are described in many monographs on Mathematical Programming, e.g., Vasil'ev [0.2], Karmanov [0.8], Ortega and Rheinboldt [0.12]. The notion of a strongly convex function was first introduced in Polyak [1.7].

**1.3.** The reader may read on the general well-posed mathematical problems in Ivanov, Vasin, and Tanana [1.4], and in Tikhonov and Arsenin [1.8]. Optimization problems are treated in Vasil'ev [0.2], Karmanov [0.8], Bank [1.12], and Fiacco [1.13]. Note that the terminology those authors have used is not always the same as in this book.

**1.4.** The idea of the gradient method is due to Cauchy (1830). The first proofs of convergence under various assumptions were constructed by Kantorovich, Veinberg, Curry, Crockett, and Chernoff in the 1940s and 1950s. Our presentation in this section basically follows the lines of Polyak [1.6]. Lemma 2 was proven by Gold'stejn and Tret'yakov [1.3]. The gradient method has been extensively treated in Ortega and Rheinboldt [0.12].

**1.5.** Linearization as a means of solving equations was used even by Isaak Newton himself, whereas the first results on convergence are credited to Fourier and Cauchy. Kantorovich made a systematic study of Newton's method in the late 1940s; see [1.5]. Various generalizations of Newton's method and the related convergence theorems are given in Ortega and Rheinboldt [0.12]. Note also that Newton's method is often referred to as the Newton-Raphson or Newton-Kantorovich method.

**1.6.** A comprehensive analysis relating to the theory versus practical implementation in computational mathematics can be found in N.S. Bakhvalov [1.1] and R.W. Hamming [1.9] (on numerical methods). We also suggest the reader should get acquainted with Fedorenko [0.19].

### Chapter 2

**2.1.** The idea of exploiting linearization to investigate asymptotic behavior of trajectories described by differentiable equations belongs to Lyapunov (the first Lyapunov method), and was elaborated by Henri Poincare, Bendixson, and many others. The first results relating to discrete processes (described

by difference equations) were obtained by Poincare and Perron. Thus the fundamental Theorem 1 was essentially proven by Perron in 1929, although it is often attributed to Ostrowski (1957). A systematic investigation of convergence of iterative procedures, based on Lyapunov's first method, was initiated by Richard Bellman, P.V. Bromberg, and Ostrowski. The theory has been treated in great detail in Ortega and Rheinboldt [0.12]—it also contains an extensive bibliography.

**2.2.** Various and numerous lemmas on numerical sequences are scattered in the literature, sometimes irrelevant to the convergence of iterative methods.

The use of scalar functions which monotonically decrease on trajectories of a process for investigating the asymptotic behavior and stability of ordinary differential equations is attributed to Lyapunov (Lyapunov's direct or second method). This method is the analytical tool most frequently used for such purposes. It has been first employed by Bellman and Bromberg to study discrete-time processes. The most complete results in this respect have been obtained by Yu.I. Lyubich and G.D. Majstrovskij [2.7]. Furthermore, in the literature on computational mathematics, proofs which involve Lyapunov's function are traditional, see e.g. Evtushenko and Zhdan [2.4]. J.B. Blum [2.12] applied Lyapunov's second method to solve stochastic problems. Thereupon it has been repeatedly used and generalized: Belen'kij, Volkonskij, et al. [2.1], K.J. Kushner [2.6], Nevelson and Khasminskij [2.8], E.A. Nurminskij [2.9] and Polyak [2.11].

The formulations and proofs of the lemmas and theorems of this section follow those in Polyak [2.11].

**2.3.** The contraction mapping principle has been traditionally used in mathematics for proving existence theorems rather than for investigating convergence—the classical proof of the existence and uniqueness of a solution of differential equations, that passes through a given point, of Charles Emile Picard. An abstract formulation of the principle was given by Stefan Banach. Various generalizations can be found in Ortega and Rheinboldt [0.12].

## Chapter 3

Methods of unconstrained minimization are emphasized in [0.2, 0.8, 0.10, 0.12, 0.13, 0.15, 0.20, 0.21, 0.22, 0.24, 0.27, 0.28, 1.2, 1.12, 3.23, 3.24]. They are also the focal point of [3.8, 3.10, 0.29, 0.30].

**3.1.** Theorem 1 on convergence of the method of steepest descent was derived by Curry in 1944. The convergence rate for quadratic functions (Theorem 2) was investigated by Kantorovich [1.5]. The procedure of regulating the step size (10) was suggested by various authors; in particular, by Armijo, Smolyak, Pshenichnyj, and Danilin [0.15]. Newton's method with regulation of step size (11) was proposed by Kantorovich in 1948. An algorithm similar to (16) for special types of problems (minimizing sums of squares of nonlinear functions) was first used by Levenberg in 1944 and later made known generally through the work of Marquardt [3.20].

Various details and more complete bibliographical references on material in this section may be found in Ortega and Rheinboldt [0.12].

**3.2.** Two-step methods for accelerating convergence of iterative procedures have been traditionally used in linear algebra [3.6, 3.11]. Method (2) was carried over to nonquadratic minimization problems in [3.7], in which convergence assertions like Theorem 1 are also derived.

The conjugate-gradient method which currently plays a significant role in the theory and practice of optimization, emerged in 1952 in works of Hestenes and Stiefel [3.17] as a technique for solving systems of linear equations with positive definite matrix. In [3.16] there is a detailed investigation of the properties of the method and of modifications of it, as well as a comparison of varied computational schemes. In 1964 Fletcher and Reeves [3.15] propounded to apply the method in the form (22) to nonquadratic problems and expressed heuristic divinations on the method's efficiency. The first proof of convergence was given by Daniel (see [3.12]). The computational scheme (24) was proposed in [0.13] and [3.9], which contain, too, proofs of convergence for variants with restart.

**3.3.** The first quasi-Newton algorithm was suggested by Davidon in 1959, and became popular due to the work of Fletcher and Powell [3.14]. The Davidon-Fletcher-Powell method is described by formulas (1), (8). Since then, there have appeared a host of new variants of quasi-Newton methods, as well as attempts to classify them and develop a unified approach [3.18], including validation of the known methods [3.2]. These questions are tackled in detail in [0.15, 0.22, 0.27, 0.28, 3.16, 3.23, 3.24] and in [3.13].

On the subject of validation of different variants of the variable metric and conjugate directions methods we refer the reader to [0.15, 0.22, 3.2].

The secant method for solving one-dimensional equations has been known for a very long time; for the two-dimensional case, as Ostrowski notes [2.10], it was suggested as early as by Gauss. In the general form, the method was described by Wolfe [3.22]. Substantiation of the method was given by Danilin [3.1]. In [3.8] some computational modifications of the method are given.

The idea of using a homogeneous model is credited to Jacobson and Oksman [3.19].

**3.4.** Various methods of finite-difference approximation and estimates of their accuracy are given in [3.4]. Methods of the form (6)-(8) were investigated in [4.2]. The method of coordinatewise descent for quadratic functions is well known in Linear Algebra as the Gauss-Seidel method [3.6, 3.11]. Rastrigin [0.16] is the pioneer in the application of random-search methods.

The simplicial method was propounded in [3.21] and generalized by many authors (see [3.3]).

The barycentric coordinate method described in Subsection 3.4.4 was constructed by Lavrov [3.5].

## Chapter 4

The effect of noise on an optimization method has not been studied thoroughly enough in the literature. The first works in this area appeared in Statistics, and were connected with stochastic approximation algorithms [4.14, 4.12]. The research on these problems, especially minimization problems, is summarized in Ermol'ev [4.2] and Kushner and Clark [4.13].

**4.1** Sources and different kinds of noise in computational problems were generally treated in works on numerical methods [1.1, 1.9]. Many examples of problems of adaptation, learning, identification, and the like, leading to a minimization of a function of the mean-risk type, are given in [4.2, 4.9, 4.11]. A classification of various kinds of noise in this section is rather unusual.

**4.2.** The case of deterministic noise was touched on in earlier works on external control [4.4] and has hardly been investigated at all in the literature on optimization methods. Problems involving random noise, as already observed, have been more amply studied [4.1, 4.2, 4.12, 4.14, 0.16, 0.18, 2.3, 2.8].

**4.3.** A comparison of the convergence rate in one- and two-step methods with noise is made in [4.6].

**4.4.** The first results like those in Theorem 1 were derived in [4.12], see also [2.3, 2.8]. In comparing the difference variant of the gradient method and the random search method, we follow the lines of [4.7]. The recursive version of the least squares method (13) is a special case of the Kalman filter.

**4.5.** The results of Subsection 4.5.1 are derived in [4.10]. Asymptotically optimal algorithms for optimization in the presence of random noise are stated and proved in [4.8]. The necessary facts from Mathematical Statistics (for example, the Cramer-Rao inequality) can be found in [4.3].

## Chapter 5

At present, there are several monographs devoted to the problem of minimizing nondifferentiable functions. Among these is Shor [5.15], a pioneer in the area of nonsmooth optimization methods. The monographs of Dem'yanov [5.2, 5.3] deal with a minimization of functions of maximum type. Ermol'ev [4.2] considers nonsmooth optimization in random noise. The question of optimal methods of convex minimization has been investigated in detail in Nemirovskij and Yudin [0.11]. Generalizations to the case of certain nonconvex smooth problems can be found in Nemirovskij [2.9] and Gupal [5.1]. Finally, there are collections of articles on this subject [5.19, 5.20, 5.31].

**5.1.** Exhaustive material on convex analysis is contained in the fundamental monograph of Rockafellar [5.13]; see also [5.11, 5.32]. The first results

on subgradients and their properties (such as Lemmas 6-12) are credited to Pshenichnyj [5.12]. Various generalizations of Lemma 7 may be found in Dem'yanov [5.2]. The properties of  $\varepsilon$ -subgradients are described in [2.9]. 5.2. Problems with a sharp minimum were thoroughly treated in [5.21].

5.3. The subgradient method (with constant step size) was first proposed by Shor in 1962 (see the references in [5.15, 5.19, 5.20]). The method of regulating the step size (4) was indicated by Ermol'ev [0.5] and B.T. Polyak [5.8]. The subgradient method in the form (7) for special problems was introduced by Eremin [5.4]. A thorough investigation of such a method was conducted in [5.9]. Variants of the method in Problem 2 are described in [5.5, 5.15]. A detailed analysis of the convergence rate of various modifications of the subgradient method and their numerical verification are given in [5.16]. Results pertaining to the  $\varepsilon$ -subgradient method are available in [2.9, 5.3]. Various generalizations of it have been suggested by Lemarechal, Wolfe, and by other authors [5.19, 5.20, 2.9]. See also surveys [5.29, 5.30].

5.4. The cutting-plane method was developed by Kelley [5.17] even before the appearance of the simpler subgradient method. Convergence results on Kelley's method may be found in [0.9]; its modifications were proposed in [5.33]. The method of Chebyshev centers is focused on in [5.6, 5.10]. The variant (6) was proposed in [5.9]. The original center-of-gravity method was developed simultaneously and independently by Levin [5.7] and Newman [5.18]. Theorem 2 in the form given here was proved by Nemirovskij and Yudin [0.11], where the authors studied in detail the question of optimal methods; they also proposed and proved the ellipsoid method (12). Shor arrived at this method from a different viewpoint. Shor also developed numerous variants of the space-dilation methods (these works are summarized in [5.15]). The form (14) for these methods was suggested by Skokov [5.14]. The ellipsoid method attracted great attention recently, see [5.23-5.28]. The new idea to replace an ellipsoid with a simplex was proposed in [5.22, 5.34].

5.5. The subgradient method in the presence of noise was studied by Ermol'ev [4.2], see also [2.9, 4.1, 5.21].

5.6. Method (3) is credited to Gupal [5.1]; he also investigated the convergence of the method.

## Chapter 6

6.1. Ill-posed problems (of which problems of unconstrained minimization with singular minimum are a special case) have received a great deal of attention of researchers in recent years. We cite for example the monographs of Tikhonov and Arsenin [1.8] and Ivanov, Vasin, and Tanana [1.4], and Lavrent'ev [6.8], focusing on solution of the linear equation  $Ax = b$  in Hilbert space when  $A$  has no bounded inverse.

↓

The result on convergence in Theorem 1 was given without proof in [6.12] and proved by Gold'stein and Tret'yakov [1.3]. Further results can be found in [6.24-6.26]. The regularization method for ill-posed problems for solving operator equations was proposed by Tikhonov in 1963 (see the reference in [1.8]). Theorem 4 was proved by Levitin and Polyak in [6.9]. Numerous results related to pseudoinverse matrices may be found in Albert [6.1]. The notion of approximate mappings was introduced by Moreau; his results are described in §31 of Rockafellar [5.13]. Method (18) was proposed by Martinet [6.19]; generalizations and a thorough analysis of it were given by Rockafellar in [6.20]. The iterative regularization method (21) was suggested by Bakuninskij and Polyak [6.2]. The idea of "regularization" involving the choice of the stopping time of an iterative process in solving linear operator equations was expressed by Bakuninskij, Lavrent'ev, as well as other authors. The most complete investigation of this approach, with constructive stopping rules, is contained in [6.6]. The regularization method in minimization problems in noise was first studied by Morozov [6.10]. Further results may be found in [0.2, 0.8, 1.8].

13h  
2P

**6.2.** The problem of global optimization is focused on in [6.15, 6.16, 6.23], in [6.21] and also in [6.3, 6.4, 6.11]. Algorithms of global optimization of Lipschitz functions have been proposed by several authors (see [6.3, 6.16]). The gully method was suggested by Gel'fand and Tsetlin [6.5]. The idea of dividing the search procedure into stages of descent, ascent and saddle point, and the method of determining the direction of the "slowest" ascent is credited to Fedorova [6.17]. Other deterministic methods of global optimization are given in [6.21].

There is an extensive literature on methods of random search. One can get acquainted with these methods in Rastrigin [0.16, 6.13]. Recent results in this area appear in the journal *Problems of Random Search* published by "Zinatne" Publishing House. Statistical models of global optimization have been developed in Motskus [6.11], Strongin [6.15], Batushchev [6.3], and Zhilinskas [6.22].

**6.3.** Constrained nonstationary optimization problems are examined in [6.7]. Originally, questions of the dependence of the solution on a parameter were studied in a theory dealing with methods of parametric extension to solve nonlinear equations (see [0.12, Secs. 7.5 and 10.4]). From a very different viewpoint nonstationary optimization problems have been investigated in works on extremal control [0.6, 4.4] and on dynamic stochastic approximation [6.18]. However, on the whole nonstationary problems have received relatively scant attention.

## Chapter 7

Considerable attention is given to the problem of minimization on a simple set in [0.2, 0.4, 0.15, 0.17, 0.24, 3.4, 4.2] and in [0.9, 0.14].

**7.1.** An extremum condition like Theorem 1 was first derived already in 1940 by Kantorovich. In the form given in this section Theorem 1 was proven by Dem'yanov and Rubinov [0.4]. Results for the nonsmooth case (Theorem 3) are credited to Pshenichnyj [5.12]. Generalized minimizing sequences and the corresponding stability condition were introduced in [0.9]. Results on "superstability" of a sharp minimum were derived in [5.21].

**7.2.** The gradient-projection method was proposed independently in [7.6, 0.4, 0.9]. The first part of Theorem 1 [a-c] was derived in [0.9]. The result on the method's finiteness in the case of a sharp minimum is contained in [5.21]. Further investigations of the gradient-projection method may be found in [7.2, 7.8]. For nonsmooth problems the subgradient-projection method (8), (9) was introduced in [5.8], and in the form (8), (10) in [5.9].

Frank and Wolfe [7.5] proposed method (11) for quadratic programming problems (i.e., with quadratic  $f(x)$  and polyhedron  $Q$ ). Therefore the method of conditional gradient is often called the Frank-Wolfe method. For the general problem (A) the conditional-gradient method was applied by Dem'yanov [0.4]. Estimates of the convergence rate and various rules for choosing the step size were proposed in [0.9, 7.4]. The method's behavior for a sharp minimum was studied in [7.4, 5.2]. An example of type (16) demonstrating the slow convergence of the method is constructed in [7.3].

Newton's method for constrained problems was introduced in [0.9].

**7.3.** Certain results similar to the material in Subsection 7.3.1 may be found in [7.7, 0.15]. The conjugate-gradient method with ~~restrictions~~ of the form  $a \leq x \leq b$  was introduced and proven in [3.9]. Regarding multistep methods for minimizing nonsmooth functions with constraints  $x \in Q$ , see references in Chapter 5.

**7.4.** Convergence of the subgradient-projection method in random noise (Theorem 2) was proved by Ermol'ev [4.1, 4.2]. The stochastic conditional-gradient method in the form (6) was proposed in [7.1]; see also [5.1]. An investigation of methods for errors satisfying condition (9) is conducted in [4.5]. A number of results pertaining to methods for solving problem (A) in random noise are available in [3.4].

constraints

✓

## Chapter 8

**8.1.** An extremum condition in the form (2) is available in most textbooks on Mathematical Analysis, e.g. [1.10].

Lyusternik's theorem (for the infinite-dimensional case) was proven in 1934; this result may be found in [8.4, 1.5, 9.1]. A proof based on the implicit-function theorem is standard in Analysis (see, e.g. [1.10]). The idea of applying penalty functions to derivation of the rule of Lagrange multipliers was propounded by several mathematicians. The proof given in this section follows that of [8.2].

The basic Lemma 3 is often called Finsler's lemma or Debreu's lemma (see e.g. [2.2] in Chap. 5). Generalizations of it to the infinite-dimensional

case and other lemmas of this section were obtained in [8.9]. The augmented Lagrangian (14) was first introduced in variational calculus in the 1930s and 1940s, to prove sufficiency conditions.

Assertions on stability like Theorem 8 are contained in [0.20]. Results for more general problems (of the type  $\min f(x, \varepsilon)$  under the constraints  $g_i(x, \varepsilon) = 0, i = 1, \dots, m$ ) may be found in [8.3, 8.5].

**8.2.** Methods for solving equality-constraint problems are examined in monographs [0.15, 0.20, 0.22] and surveys [0.14, 8.11].

The method of linearization was proposed at a conceptual level by various authors. As a precise algorithm, it was introduced by Pshenichnyj [8.10]; these results are described in detail in [8.20, 0.15]. Our version of the method differs from the one in [0.15] by a simpler rule for choosing the step size.

The Lagrange-multiplier methods (6), (8) were systematically studied by Arrow, Hurwicz, and Uzawa [0.23] mainly for convex problems; problems with equality constraints are considered in Chapter 11 thereof. Theorem 2 and the convergence result for method (8) were proven by this author [8.6].

The augmented Lagrangian method in the form (11) was independently suggested in [8.13-8.15]. Theorems 4 and 5 on its convergence and rate of convergence were proved by Tret'yakov and this author [8.9]; see also [8.11]. A variant of method (9) for linear constraints was proposed by Antipin [8.1], and for the general case in [8.8]. New forms of the augmented Lagrangian are suggested in [8.18, 8.19].

The penalty-function method was historically the first method of constrained minimization; Courant introduced it in 1943, and Butler and Martin were the first to prove it in 1962. An estimate of the convergence rate (Theorem 7) was derived in [8.7]. Let us also mention [0.9, 0.20, 0.22, 0.31], where further references may be found, too.

The reduced-gradient method has been used in engineering for a long time. It was investigated by Wolfe [8.17]; see also [0.22].

Newton's method (19) is a straightforward approach to the problem involving the rule of Lagrange multipliers. Conditions for its convergence (Theorem 9) were derived in [8.6].

Robinson [8.16] reduced the problem of constrained minimization to a sequence of problems with linear constraints of the form (25). He is also credited with a result on quadratic convergence of the method. Methods of the form (26) were investigated in [0.22].

**8.3.** The behavior of optimization methods with equality constraints in random noise was thoroughly studied in [8.8].

## Chapter 9

**9.1.** The theory of convex programming is treated in several texts and monographs [9.9, 9.26, 0.7, 0.8, 0.20, 0.23, 0.24, 2.5, 5.11, 5.12].

As has been observed, the most complete textbook on Convex Analysis is Rockafellar [5.13], where one may find proofs of Lemmas 1-4. The fundamental Lemma 4 is credited to Moreau and Rockafellar (see the references in [5.13]). Lemma 5 was derived in the infinite-dimensional case by Dubovitskij and Milyutin [9.7]. This result underpins their widely used technique for analyzing general extremal problems (for this see also [9.3, 9.7, 9.12, 9.28]). Farkas' lemma, or assertions equivalent to it, have been known since the 1930s, and it is the basic tool in finite-dimensional theory of linear and convex programming. The basic Theorems 1 and 3 were proven by Kuhn and Tucker in 1950; they also emphasized the major role of the regularity condition, which was earlier introduced by Slater.

The dual problem was originally constructed for a problem of linear programming (see the references in Chap. 10). Later, this was done for quadratic programming, although various authors introduced the notion of a dual problem in different ways. Duality theory was in even greater chaos with respect to general problems of convex programming. In a number of monographs and textbooks one can find, for instance, a definition of the dual problem in which both  $x$  and  $y$  appear. The duality theory presented in Subsection 9.1.3 was developed by Gold'stein; in [9.4] much more general results are given (see also [9.9]). The most recent approach to duality involving conjugate functions and perturbations of general form is delineated in Rockafellar [5.13] and Ioffe and Tikhomirov [9.12].

General formulations of the stability problem may be found in [9.4, 9.8, 1.12, 1.13, 8.3, 8.5].

**9.2.** Theorem 1, one of the basic results in the theory of nonlinear programming, is a direct generalization of the rule of Lagrange multipliers. For a long time, together with Theorem 1 of Section 9.1, it was called the Kuhn-Tucker theorem. The fact is that this result was obtained earlier by John. However, as recently came to light, it was Karush who found extremum conditions for smooth problems of first and second orders as early as in 1939. A complete and enlightening history of the discovery of the "Karush-John-Kuhn-Tucker theorem" is described in Kuhn's article [9.25]; the Karush manuscript is included there as an application. Various extremum conditions and regularity conditions, and relationships between them, are investigated in utmost detail in Mangasarian [9.26] (see also [0.20]). The most recent schemes for obtaining general extremum conditions for both smooth and non-smooth problems in the infinite-dimensional case are being developed very intensively and have reached a high level of abstraction (see [9.1, 9.3, 9.7, 9.12, 9.28]).

Stability conditions like Theorem 9 were studied by Fiacco and MacCormick [0.20].

**9.3.** The works of Zoutendijk from the late 1950s and early 1960s are summarized in his monograph [0.7], which contains a general construction of methods of feasible directions, and numerous modifications and specific

forms of general algorithms for special cases. Independently of Zoutendijk, similar schemes were proposed and investigated by Zhukovskij, R.A. Polyak and Primak [9.11, 9.10], among others; see survey [0.14]. One of the first realizations of the method of feasible directions was Rosen's gradient-projection method [9.31]. Proofs of convergence may be found in [0.7, 9.31, 9.11].

The linearization method, with proof, was first proposed by Pshenichnyj in [8.10]. His results are presented in detail in [0.15, 8.20]. Theorem 1 pertaining to the convergence of the method with constant  $\gamma$  is credited to Antipin [9.2].

The method of Lagrange multipliers was one of the first numerical methods of convex programming. It is discussed in Arrow, Hurwicz, and Uzawa [0.23]. These authors mainly examine continuous variants of the method (the trajectory of motion is described rather by differential equations than by difference equations), and one chapter in [0.23], written by Uzawa, is concerned with the discrete method. Also, they investigate method (13) and state that if  $f(x)$  is strictly convex, then for every initial point and any  $\epsilon > 0$  there is a sufficiently small  $\gamma > 0$  such that the method leads into an  $\epsilon$ -neighborhood of the solution. Unfortunately there is a major error in the proof of this assertion (Theorem 1 in Chap. 10 of [0.23]). Taking this approach it is possible to prove [0.6] only that there is a subsequence  $x^*$  falling into a neighborhood of the point  $x^*$ . The convergence of a method of the form (13) is proven differently in [9.14]. Method (16) was also proposed in [0.23], with proof of the method's convergence in particular cases.

The augmented Lagrangians for convex programming problems were introduced and studied by Bertsekas, Wierzbicki, Gold'stejn, Rockafellar, and Tret'yakov [9.6, 9.18, 9.20, 9.30, 9.32]. Iterative methods of the type (20) for linear constraints were studied by Antipin [8.1]. A result which includes Theorem 2 as a special case was derived by Majstrovskij [9.15], also see [9.5]. Theorem 3 was proven by Tret'yakov [9.18] and Rockafellar [9.30].

The method of simultaneous solution of the primal and dual problem (25) and a dual scheme of a linearization method similar to it were investigated in [9.16, 9.17].

The monographs [0.20, 9.22] and the survey [9.19] deal with the methods of penalties, barriers and the like. Estimates of the convergence rate for method (26) are given in [9.8]. Results similar to Theorems 4, 5 may be found in [0.20]. The penalty-shift method was introduced by Wierzbicki [9.32]; it was later proven to coincide with the augmented-Lagrangian method.

The method of choosing  $f^*$  was proposed by Morrison [9.27]. Other ways of updating the  $f_k$  and results on convergence may be found in [9.13, 9.24].

A generalization of the subgradient method to convex programming problems was proposed in [5.8]. The equivalence of the problem of convex programming to a problem of unconstrained minimization with nonsmooth penalties (36) was observed by a number of authors [9.8, 9.29]. One can get acquainted with a generalization of the cutting-plane method and some other methods mentioned in Subsection 9.3.5 in the literature referred to in Section 5.4.

**9.4.** Methods for solving a general problem of nonlinear programming (1) are examined in monographs [0.15, 0.20, 0.22] and survey [0.14]. Results like Theorems 1, 2 on convergence of the linearization method (for a somewhat different rule of choosing  $\gamma$ ) were derived by Robinson in [8.16]. Computational schemes of quasi-Newton methods are given in [0.22, 9.21, 9.23].

## Chapter 10

There are many texts and monographs devoted mainly to the problem of linear programming [0.7, 0.32, 9.10, 10.1-10.5, 10.7, 10.11, 10.14, 10.16, 10.17]. The focus of attention is on the algorithmic implementation of the simplex method and its generalizations.

**10.1.** The honor of formulating the problem of linear programming belongs to Kantorovich [10.6]. Since then an enormous number of investigations have been devoted to this problem (see e.g. the bibliography in [10.2, 10.5]). Results like Theorem 1 were derived by Caratheodory and Weyl. Contributors to the development of optimality conditions are also Kantorovich, Dantzig, von Neumann, Gale, Kuhn, Tucker (see [10.10]). Theorem 12 for special cases was formulated in [0.23, 10.12] and for the general case in [5.21].

**10.2.** The simplex method was proposed by Dantzig. A detailed presentation of the computational scheme and modifications may be found in [0.7, 0.32, 9.10, 10.1-10.3, 10.5, 10.11, 10.14, 10.16, 10.17].

**10.3.** The finiteness of the augmented Lagrangian method was shown by this author and Tret'yakov [10.12]. Solution of linear programming problems by reducing them to unconstrained minimization of piecewise linear functions was first suggested by Shor in the early 1960s (see [5.15]). The work of Khachiyan [10.15] mentioned in this section was published in 1979 but has already exerted a great influence on the views shared by specialists in linear programming; see e.g. [5.23, 5.27, 5.28, 5.34]. Results like Theorem 5 were derived by Antipin [8.1] and Majstrovkij [9.15]. The "fictitious play" method for solving matrix games was proposed by Brown (see [10.4]). A general formulation and proof of this method were given by Volkonskij, et al. [2.1]. Besides the methods described in the text, there are also many other iterative methods of linear programming [0.7, 0.23, 10.4, 10.8, 10.13, 10.22]. A large-dimensional numerical experiment for one of these methods appears in Fedorenko [0.19]. There are new ideas in constructing iterative methods, which seem to be very promising [10.21, 10.23].

Formula (17) for the total number of polyhedra generated by  $m$  hyperplanes in  $\mathbb{R}^n$  is due to Schl  fli (1852). Intensive investigations concerning the average number of steps of the simplex method are summarized in [10.19, 10.20, 10.24, 10.25].

**10.4.** Various quadratic programming algorithms (developed in the early 1960s) are described in [0.7, 0.26, 10.9].

## Chapter 11

Applied optimization problems are the focus of monographs [11.33, 11.36] and of survey [11.34], and many examples of such problems may be found in [0.1, 0.16, 0.21, 0.25, 0.31].

**11.1.** The statement and methods for solving the classical problem of parameter estimation are described in many textbooks on Mathematical Statistics; see e.g. [4.3]. Other examples of optimization problems in statistics are given in [11.35, 11.42].

A survey of methods for regression problems is given in [11.3, 11.8, 11.37]. Results of calculations for the nonlinear least squares method are given in [11.22]. The concept of robust estimation was introduced by Huber [11.40], see also [11.13]. Method (18) is described in [11.18], method (20) in [11.32]: they are verified in [11.38] and [11.43], respectively. General statistical approaches to the problem of reconstructing relationships from experimental data are investigated by Vapnik [11.5].

Adaptive methods have gained extensive popularity through works of Tsyplkin [4.9]. The recursive variant of the maximum likelihood method (22), (23) was proposed by Sakrison (see [2.8]). Recursive estimates for regression problems were investigated by Tsyplkin and this author [11.23].

Problems of choosing models from experimental data were analyzed in [11.3, 11.29, 11.37]. One can familiarize oneself with estimating parameters in dynamic models in [11.1, 11.41], and with problems of optimal experiment design in [11.19, 11.27].

**11.2.** Examples of applied problems of optimal design are contained in [11.6, 11.7, 11.20, 11.30, 6.3, 6.11]. Geometric programming problems are dealt with in [11.10]: the bibliography on this topic may be found in [11.39].

A detailed investigation of various problems of allocating resources is contained in [11.9]. Dynamic programming as developed by Bellman [11.4] (see also [11.2]) is a special method for solving optimization problems which have a dynamic structure. The ideas of the method can also be used for certain other problems—for example, problem (8).

Optimization problems in economics were first investigated by Kantorovich [10.6]; numerous examples of statements of economic problems are contained in [11.15, 11.21, 10.4-10.6, 10.16]. Decomposition methods are discussed in [11.17, 11.26].

Various statements of stochastic programming problems and methods for solving them are considered in monographs of Ermol'ev [4.2, 11.12] and Yudin [4.11]; also see [4.15].

Minimax problems were intensely investigated by Dem'yanov, Fedorov and others [5.2, 5.3, 11.28, 2.9]. Formulations of optimization problems under uncertainty conditions, other than minimax problems, were examined by Germeyer [11.8].

One can get acquainted with extremal control problems in [0.16, 11.14]. Extensive literature is devoted to theory and numerical methods for solving discrete optimal control problems; let us cite only [11.2, 11.4, 11.24, 11.25, 0.2, 0.10, 0.13, 0.19].

**11.3.** The connection of approximation theory with optimization theory is discussed in [11.16, 5.12]. Some geometric extremum problems are given in [9.1, 9.12].

A systematic study of interrelations between the theory of electric circuits and mathematical programming was undertaken by Dennis [11.11].

Razumikhin [10.13] examines the application of variational principles of physics to constructing numerical optimization methods. A number of examples of variational problems from mechanics are given in the monographs of Chernous'ko and Banichuk [11.31].

## Chapter 12

**12.1.** Techniques and procedures for formulating and solving optimization problems are rarely touched upon in the literature on optimization. The general concept of interactions between mathematician/computer analyst and practitioner/user is well explained in Bakhvalov [1.1] dealing with numerical methods. The process of solving real optimization problems (principally optimal control problems) is illustrated with a great deal of source material in Fedorenko [0.19]. In this connection, books on operations research [11.8, 12.6] are also useful.

**12.2.** Program specifications for nonlinear programming are formulated in [12.6]. Software texts for unconstrained minimization of smooth functions are available in [0.21, 12.12, 12.13]; for nonsmooth ones in [11.6]; for linear programming in [10.14, 12.12-12.15]; for quadratic programming in [12.14]; for nonlinear programming in [0.21, 12.12, 12.20] and also in publications cited throughout the text. A survey of available programs for linear programming is given in [12.4] (the status as of 1969) and in Chapter 16 of [12.6] (the current state) (see also [12.7, 12.8, 12.15]). The existing man-machine optimization systems are given in Chapter 15 of [12.6], a concrete man-machine system is described in [0.10]. A package of optimization programs developed under this author's direction is described in [12.1, 12.2]. A complete list of nonlinear programming programs available in the West is contained in [12.16].

**12.3.** A discussion of algorithm comparison criteria is available in [0.21]; there is also given a special timer-program for comparing the speed of various computers. Standardized requirements on publications for comparing numerical optimization algorithms prepared by a special commission are formulated in [12.10]; see also [12.19]. The important problem, "How to compare complexity of calculation of a function and its derivative?" is discussed in [12.18].

The first systematic collection of test problems and investigations of various methods with these problems was Colville's work [12.9]. Since then many of the tests appearing in [12.9] have become classical (e.g. Rosenbrock's function) and new programs and methods of optimization are analyzed using them. A large collection of unconstrained and constrained extremum test problems are given in Himmelblau [0.21] and in [12.17]: they contain results of numerical experiments in verifying many methods by means of these tests. Tests for unconstrained minimization of nonsmooth functions are given in [5.20]. A large experiment of solution of nonlinear programming test problems is described in Skokov's article [12.15]. Tests pertaining to geometric programming problems are collected in [12.11]. An important example of a linear programming problem which is difficult for the simplex method is given in [10.18], and problem 11 is built around it. On the whole, it may be said that current work in developing tests of varied complexity for different classes of problems, systematizing them and verifying algorithms with these tests, is still far from being complete.

## REFERENCES

### FUNDAMENTALS: TEXTBOOKS, MONOGRAPHS, SURVEYS

- 0.1. Aoki, Masanao. *Introduction to Optimization Techniques; Fundamentals and Applications of Nonlinear Programming*. New York: Macmillan; London: Collier-Macmillan, 1971.
- 0.2. Vasil'ev, F.P. *Lektsii po Metodam Resheniya Ekstremal'nykh Zadach* (Lectures on the Methods for Solving Extremal Problems). Moscow: Izd. MGU, 1974.
- 0.3. Gabasov, P., and Kirillova, F.M. *Methods of Optimization*. New York: Optimization Software, Inc., Publications Division. Forthcoming. (English transl.)
- 0.4. Dem'yanov, V.F., and Rubinov, A.M. *Approximate Methods of Solving Extremal Problems*. New York: American Elsevier, 1970. (English transl.)
- 0.5. Ermol'ev, Yu.M. "Methods for Solving Nonlinear Extremal Problems" (in Russian). *Kibernetika* 4 (1966): 1-17.
- 0.6. Zangwill, Willard I. *Nonlinear Programming; a Unified Approach*. Englewood Cliffs, N.J.: Prentice-Hall, 1969.
- 0.7. Zoutendijk, G. *Methods of Feasible Directions; a Study of Linear and Non-linear Programming*. Amsterdam: Elsevier, 1960.
- 0.8. Karmanov, V.G. *Matematicheskoe Programmirovaniye* (Mathematical Programming). Moscow: Nauka, 1975.
- 0.9. Levitin, E.S., and Polyak, B.T. "Constrained Minimization Methods." *U.S.S.R. Comput. Maths. Math. Phys.* 6, 5 (1966): 1-50. (English transl.)
- 0.10. Moiseev, N.N., Ivanilov, Yu.P., and Stolyarova, E.M. *Metody Optimizatsii* (Optimization Methods). Moscow: Nauka, 1978.
- △ 0.11. Nemirovskij, A.S., and Yudin, D.B. *Slozhnost' Zadach i Effektivnost' Metodov Optimizatsii* (Complexity of Problems and Effectiveness of Optimization Methods). Moscow: Nauka, 1980.
- 0.12. Ortega, J.M., and Rheinboldt, W.C. *Iterative Solution of Nonlinear Equations in Several Variables*. New York: Academic Press, 1970.
- △ 0.13. Polak, E. *Computational Methods in Optimization; a Unified Approach*. New York: Academic Press, 1971.
- △ 0.14. Polyak, B.T. "Methods of Constrained Minimization" (in Russian). In *Itogi Nauki i Tekhniki. Matematicheskiy Analiz* (Moscow, VINITI), vol. 12 (1974): 147-97.
- 0.15. Pshenichnyj, B.M., and Danilin, Yu.M. *Numerical Methods in Extremal Problems*. Moscow: Mir, 1978. (English transl.)
- 0.16. Rastrigin, L.A. *Statisticheskie Metody Poiska Ekstremuma* (Statistical Extremum Seeking Methods). Moscow: Nauka, 1968.

- 0.17.** Céa, Jean. *Lectures on Optimization Theory and Algorithms*. Bombay, Tata Institute for Fundamental Research, Bombay. Berlin Heidelberg New York: Springer-Verlag, 1978.
- 0.18.** Wilde, D.J. *Optimum Seeking Methods*. Englewood Cliffs, N.J.: Prentice-Hall, 1964.
- 0.19.** Fedorenko, R.P. *Priblizhennoe Reshenie Zadach Optimal'nogo Upravleniya* (Approximate Solution of Optimal Control Problems). Moscow: Nauka, 1978.
- 0.20.** Fiacco, Anthony V., and McCormick, Garth P. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. New York: John Wiley & Sons Publ. Co., 1968.
- 0.21.** Himmelblau, David M. *Applied Nonlinear Programming*. New York: McGraw-Hill, 1972.
- 0.22.** Gill, P., and Murray, W., eds. *Numerical Methods for Constrained Optimization*. London New York: Academic Press, 1974.
- 0.23.** Arrow, Kenneth J., Hurwicz, Leonid, and Uzawa, Hirofumi. *Studies in Linear and Non-linear Programming*. Stanford, California: Stanford University Press, 1958.
- 0.24.** Auslender, A. *Optimisation: méthodes numériques*. Paris: Masson, 1976.
- 0.25.** Bazaraa, M.S., and Shetty, C.M. *Nonlinear Programming*. New York: John Wiley & Sons Publ. Co., 1979.
- 0.26.** Blum, E., and Oettli, W. *Mathematische Optimierung*. Berlin Heidelberg: Springer-Verlag, 1975.
- 0.27.** Fletcher, R. *Practical Methods of Optimization*. Vols. 1, 2. Chichester, N.Y.: John Wiley & Sons Publ. Co., 1980-81.
- 0.28.** Gill, P.E., Murray, W., and Wright, M.H. *Practical Optimization*. London: Academic Press, 1981.
- 0.29.** Powell, M.J.D., ed. *Nonlinear Programming*. London: Academic Press, 1982.
- 0.30.** Bachem, A., Grotschel, M., and Korte, B., eds. *Mathematical Programming. The State-of-the-Art*. Berlin Heidelberg New York Tokyo: Springer-Verlag, 1983.
- 0.31.** McCormick, G.P. *Nonlinear Programming—Theory, Algorithms and Applications*. New York: John Wiley & Sons Publ. Co., 1983.
- 0.32.** Shapiro, J.E. *Mathematical Programming Structures and Algorithms*. New York: John Wiley & Sons Publ. Co., 1979.

## Chapter 1

(see also 0.1-0.3, 0.8, 0.10, 0.12, 0.13, 0.15, 0.17, 0.21)

- 1.1.** Bakhvalov, N.S. *Chislennye Metody* (Numerical Methods). Moscow: Nauka, 1973.
- 1.2.** Vainberg (Vainberg), M.M. *Variational Method and the Method of Monotone Operators in the Theory of Nonlinear Equations*. New York: John Wiley & Sons Publ. Co., 1973. (English transl.)

- 1.3. Gol'stejn, E.G., and Tret'yakov, N.V. "The Gradient Method of Minimization and Algorithms of Convex Programming Related to Augmented Lagrangians" (in Russian). *Ekonomika i Matem. Methody* 11, 4 (1975): 730-42.
- 1.4. Ivanov, V.K., Vasin, V.V., and Tanana, V.P. *Teoriya Linejnykh Nekorrektnykh Zadach i ee Prilozheniya* (Theory of Linear Ill-posed Problems and its Applications). Moscow: Nauka, 1978.
- 1.5. Kantorovich, L.V., and Akilov, G.P. *Functional Analysis in Normed Spaces*. New York: Macmillan, 1964. (English transl.)
- 1.6. Polyak, B.T. "Gradient Methods for the Minimization of Functionals." *U.S.S.R. Comput. Maths. Math. Phys.* 3, 4 (1963): 864-78. (English transl.)
- 1.7. ——— (Poljak). "Existence Theorems and Convergence of Minimizing Sequences in Extremum Problems with Restrictions." *Soviet Math. Dokl.*, vol. 7, no. 1 (1966): 72-75. (English transl.)
- 1.8. Tikhonov, A.N., and Arsenin, V.Ya. *Méthodes de resolution de problèmes mal posés*. Moscow: Mir, 1976. (French transl.)
- 1.9. Hamming, Richard W. *Numerical Methods for Scientists and Engineers*. New York: McGraw-Hill, 1973. (2d ed. 1973).
- 1.10. Shilov, G.E. *Matematicheskij Analiz. Funktsii Neskolkikh Veshchestvennykh Peremennykh* (Mathematical Analysis. Functions of Several Real Variables). Pts. 1, 2. Moscow: Nauka, 1972.
- 1.11. A n o n y m o u s. "A New Algorithm for Optimization." *Math. Programming* 3, 2 (1972): 124-8.
- 1.12. Bank, B., Guddat, J., Klatte, D., Kummer, B., and Tammer, K. *Nonlinear Parametric Optimization*. Berlin: Akademie-Verlag, 1982.
- 1.13. Fiacco, A.V. *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*. New York: Academic Press, 1983.
- 1.14. Goldstein, A.A. *Constructive Real Analysis*. New York: Harper & Row, 1967.

## Chapter 2

(see also 0.6, 0.12, 0.13)

- 2.1. Belen'kij, V.Z., Volkonskij, V.A., et al. *Iterativnye Metody v Teorii Igr i Programmirovaniu* (Iterative Methods in the Theory of Games and Programming). Moscow: Nauka, 1974.
- 2.2. Bellman, R. *Introduction to Matrix Analysis*. New York: McGraw-Hill, 1960.
- 2.3. Wasan, M. *Stochastic Approximation*. Cambridge Tracts in Math. and Math. Physics, vol. 58. Cambridge, London: Cambridge University Press, 1969.
- 2.4. Evtushenko, Yu.G., and Zhadan, V.G. "Application of the Method of Lyapunov Functions to the Study of the Convergence of Numerical Methods." *U.S.S.R. Comput. Maths. Math. Phys.* 15, 1 (1975): 96-108. (English transl.)
- 2.5. Karlin, Samuel. *Mathematical Methods and Theory in Games, Programming, and Economics*. Reading, Mass.: Addison-Wesley Pub. Co., 1959.

- 2.6.** Kushner, Harold J. *Stochastic Stability and Control*. New York: Academic Press, 1967.
- 2.7.** Lyubich, Yu.I., and Majstrovskij, G.D. "A General Theory of Relaxation Processes for Convex Functionals." *Russian Math. Surveys* 25, 1 (1970): 57-112. (English transl.)
- 2.8.** Nevel'son, M.B., and Khas'minskij, R.Z. *Stochastic Approximation and Recursive Estimation*. Providence, R.I.: American Math. Society, 1973. (English transl.)
- 2.9.** Nurminskij, E.A. *Chislennye Metody Resheniya Determinirovannykh i Stokhasticheskikh Minimaksnykh Zadach* (Numerical Methods for Solving Deterministic and Stochastic Minimax Problems). Kiev: Naukova Dumka, 1979.
- 2.10.** Ostrowski, Alexander M. *Solution of Equations and Systems of Equations*, 2d ed., New York: Academic Press, 1966; 3d ed. *Solution of Equations in Euclidean and Banach Spaces*, 1973.
- △ **2.11.** Polyak, B.T. "Convergence and the Rate of Convergence of Iterative Stochastic Algorithms. I. General Case" (in Russian). *Avtomatika i Telemechanika* 12 (1976): 83-94.
- 2.12.** Blum, J.B. "Multidimensional Stochastic Approximation Method." *Ann. Math. Statist.* 25, 4 (1954): 737-44.

### Chapter 3

(see also 0.1-0.3, 0.8, 0.10, 0.11-0.13, 0.15-0.17, 0.20, 0.22)

- 3.1.** Danilin, Yu.M. "On a Class of Minimization Algorithms with Over-linear Convergence." *U.S.S.R. Comput. Maths. Math. Phys.* 14, 3 (1974): 59-71. (English transl.)
- 3.2.** ——. "Convergence Rate of Methods of Conjugate Directions." *Cybernetics*, vol. 13, no. 6 (1977): 892-902. (English transl.)
- 3.3.** Dambrauskas, A.P. *Simpleksnyj Poisk* (Simplex Search). Moscow: Energiya, 1979.
- 3.4.** Katkovnik, V.Ya. *Linejnye Otsenki i Stokhasticheskie Zadachi Optimizatsii* (Linear Estimators and Stochastic Problems of Optimization). Moscow: Nauka, 1976.
- 3.5.** Lavrov, S.S. "Use of Barycentric Coordinates for Solving Certain Computational Problems." *U.S.S.R. Comput. Maths. Math. Phys.* 4, 5 (1964): 157-66. (English transl.)
- 3.6.** Marchuk, G.I. *Methods of Numerical Mathematics*. Berlin Heidelberg New York: Springer-Verlag, 1975. (English transl.) (2d ed. in Russian: Moscow: Nauka, 1980.)
- 3.7.** Polyak, B.T. "Some Methods of Speeding up the Convergence of Iteration Methods." *U.S.S.R. Comput. Maths. Math. Phys.* 4, 5 (1964): 1-17. (English transl.)

- 3.8.** ——. "Methods for Minimizing Functions of Many Variables: a Survey" (in Russian). *Ekonomika i Matemat. Metody* 3, 6 (1967): 881-902.
- 3.9.** ——. "The Conjugate Gradient Method in Extremal Problems." *U.S.S.R. Comput. Maths. Math. Phys.* 9, 4 (1969): 94-112. (English transl.)
- 3.10.** Saul'ev, V.K., and Samojlova, I.I. "Approximate Methods of Unconstrained Optimization of Functions of Many Variables." In *Itogi Nauki i Tekhniki. Matematicheskij Analiz* (Moscow, VINITI), vol. 11 (1973): 91-128.
- 3.11.** Faddeev, D.K., and Faddeeva, V.N. *Computational Methods of Linear Algebra*. San Francisco: Freeman, 1963. (English transl.)
- 3.12.** Daniel, J.W. *The Approximate Minimization of Functionals*. Englewood Cliffs, N.J.: Prentice-Hall, 1971.
- 3.13.** Dennis, J.E., and More, J.J. "Quasi-Newton Methods: Motivation and Theory." *SIAM Review* 19, 1 (1977): 46-89.
- 3.14.** Fletcher, R., and Powell, M.J. "A Rapidly Convergent Descent Method for Minimization." *Comput. J.* 6, 2 (1963): 163-8.
- 3.15.** Fletcher, R., and Reeves, C.M. "Function Minimization by Conjugate Gradients." *Comput. J.* 7, 2 (1964): 149-54.
- 3.16.** Hestenes, M.R. *Conjugate Direction Methods in Optimization*. New York Berlin Heidelberg Tokyo: Springer-Verlag, 1980.
- 3.17.** Hestenes, M.R., and Stiefel, E. "Methods of Conjugate Gradients for Solving Linear Systems." *J. Res. Nat. Bur. Stand. USA* 49, 6 (1952): 409-36.
- 3.18.** Huang, H.G. "Unified Approach to Quadratically Convergent Algorithms for Function Minimization." *J. Optim. Theory Appl.* 5, 6 (1970): 405-23.
- 3.19.** Jacobson, D., and Oksman, W. "An Algorithm That Minimizes Homogeneous Function of  $n$  Variables in  $n+2$  Iterations and Rapidly Minimizes General Functions." *J. Math. Anal. Appl.* 38, 3 (1972): 535-52.
- 3.20.** Marquardt, D.W. "An Algorithm for Least Squares Estimation of Nonlinear Parameters." *J. SIAM* 11, 2 (1963): 431-41.
- 3.21.** Spendley, W., Hext, G.R., and Hinsworth, F.R. "Sequential Application of Simplex Design in Optimization and Evolutionary Operation." *Technometrics* 4, 4 (1962): 441-61.
- 3.22.** Wolfe, P. "The Secant Method for Simultaneous Nonlinear Equations." *Comm. ACM* 2, 1 (1959): 12-3.
- 3.23.** Brent, M.P. *Algorithms for Minimization Without Derivatives*. Englewood Cliffs, N.J.: Prentice-Hall, 1973.
- 3.24.** Dennis, J.E., Jr., and Schnabel, R.B. *Numerical Methods for Nonlinear Equations and Unconstrained Optimization*. Englewood Cliffs, N.J.: Prentice-Hall, 1983.

## Chapter 4

(see also 0.18, 1.1, 1.3, 2.1, 2.7-2.9, 3.4)



- 4.1.** Ermol'ev, Yu.M. "On the Method of Generalized Stochastic Gradients and Stochastic Quasi-Fejér Sequences" (in Russian). *Kibernetika* 2 (1969): 73-83.



English translation exists

- 4.2. ——. *Metody Stokhasticheskogo Programmirovaniya* (Stochastic Programming Methods). Moscow: Nauka, 1976.
- 4.3. Cramer, Harald. *Mathematical Methods of Statistics*. Princeton, New Jersey: Princeton University Press, 1963.
- 4.4. Pervozvanskij, A.A. *Sluchajnye Protsessy v Nelinejnykh Avtomaticheskikh Sistemakh* (Random Processes in Nonlinear Automation Systems). Moscow: Fizmatgiz, 1962.
- 4.5. Polyak, B.T. “The Convergence of the Methods of Feasible Directions in Extremal Problems.” *U.S.S.R. Comput. Maths. Math. Phys.* 11, 4 (1971): 53-70. (English transl.)
- 4.6. ——. “A Comparison of the Rate of Convergence of One-step- and Multi-step Algorithms of Optimization in Noise” (in Russian). *Tekhnicheskaya Kibernetika* 1 (1977): 9-12.
- 4.7. ——. “On the Comparison Between The Gradient Method and the Random Search Method” (in Russian). *Avtomatika i Vychislitel'naya Tekhnika* 3 (1977): 194-97.
- 4.8. Polyak, B.T., and Tsyplkin, Ya.Z. “Optimal Pseudogradient Algorithms of Adaptation” (in Russian). *Doklady Akademii Nauk SSSR*, Ser. Kibernetika i Teoriya Regulirovaniya, vol. 250, no. 5 (1980): 1084-87.
- 4.9. Tsyplkin, Ya.Z. *Adaptatsiya i Obuchenie v Avtomaticheskikh Sistemakh* (Adaptation and Learning Techniques in Automation Systems). Moscow: Nauka, 1968.
- 4.10. Tsyplkin, Ya.Z., and Polyak, B.T. “An Attainable Accuracy of Adaptation Algorithms” (in Russian). *Doklady Akademii Nauk SSSR*, Ser. Kibernetika i Teoriya Regulirovaniya, vol. 218, no. 3 (1974): 532-35.
- 4.11. Yudin, D.B. *Zadachi i Metody Stokhasticheskogo Programmirovaniya* (Problems and Methods of Stochastic Programming). Moscow: Sovetskoe Radio, 1979.
- 4.12. Kiefer, J., and Wolfowitz, J. “Stochastic Estimation of the Maximum of a Regression Function.” *Ann. Math. Statist.* 23, 3 (1952): 462-66.
- 4.13. Kushner, H.J., and Clark, D.S. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. New York Berlin Heidelberg: Springer-Verlag, 1978.
- 4.14. Robbins, H., and Monro, S. “A Stochastic Approximation Method.” *Ann. Math. Statist.* 22, 3 (1951): 400-407.
- 4.15. Wets, R.J.-B. “Stochastic Programming: Solution Techniques and Approximation Schemes.” In *Mathematical Programming. The State-of-the-Art*, ed. A. Bachem, M. Grötschel, and B. Korte, 566-603. Berlin Heidelberg New York Tokyo: Springer-Verlag, 1983.

## Chapter 5

(see also 0.2, 0.5, 0.8, 0.9, 0.11, 0.18, 2.9, 3.4, 4.1, 4.2, 4.11)

### 5.1. Gupal, A.M. *Stokhasticheskie Methody Resheniya Negladkikh Ekstre-*

- mal'nykh Zadach* (Stochastic Methods for Solving Nonsmooth Extremal Problems). Kiev: Naukova Dumka, 1979.
- 5.2. Dem'yanov, V.F. *Minimaks: Differentsiruemost' po Napravleniyam* (Minimax: Directional Differentiability). Leningrad: Izd. LGU, 1974.
- 5.3. Dem'yanov (Dem'janov), V.F., and Malozemov, V.N. *Introduction to Minimax*. New York: John Wiley & Sons Publ. Co., 1974. (English transl.)
- 5.4. Eremin, I.I. "The Relaxation Method of Solving Systems of Inequalities with Convex Functions in the Left Sides." *Soviet Math. Doklady*, vol. 6, no. 1 (1965): 219-23. (English transl.)
- 5.5. Ermol'ev, Yu.M., and Shor, N.Z. "On Minimization of Nondifferentiable Functions" (in Russian). *Kibernetika* 1 (1967): 101-102.
- 5.6. Zukhovitskij (Zuhovichkii), S.I., and M.E. Primak. "On the Convergence of the Method of Chebyshev Centers and the Method for Centered Sections for Solving a Convex Programming Problem." *Soviet Math. Dokl.*, vol. 16, no. 3 (1975): 615-18. (English transl.)
- 5.7. Levin, A.Yu. "On an Algorithm for Minimization of Convex Functions." *Soviet Math. Dokl.*, vol. 6, no. 1: (1965): 286-90. (English transl.)
- 5.8. Polyak (Poljak), B.T. "A General Method of Solving Extremum Problems." *Soviet Math. Dokl.*, vol. 8, no. 3 (1967): 593-97. (English transl.)
- 5.9. —. "Minimization of Nonsmooth Functionals." *U.S.S.R. Comput. Maths. Math. Phys.* 9, 3 (1969): 14-29. (English transl.)
- 5.10. Primak, M.E. "Convergence of Modified Method of Chebyshev Centers for Solving Problems of Convex Programming." *Cybernetics*, vol. 13, no. 5 (1977): 738-41. (English transl.)
- 5.11. Pshenichnyj, B.N. *Convex Analysis and Extremal Problems* (in Russian). Moscow: Nauka, 1980.
- 5.12. — (Psenicnyi). *Necessary Conditions for an Extremum*. Transl. ed. by Lucien W. Neustadt. New York: Marcel Dekker, 1971. (English transl.)
- 5.13. Rockafellar, R.T. *Convex Analysis*. Princeton Math. Series 28, Princeton, New Jersey: Princeton University Press, 1970.
- 5.14. Skokov, V.A. "Note on Minimization Methods Employing Space Stretching." *Cybernetics*, vol. 10, vol. 4 (1974): 689-92. (English transl.)
- 5.15. Shor, N.Z. *Minimization Methods for Nondifferentiable Functions*. Berlin Heidelberg New York Tokyo: Springer-Verlag, 1985. (English transl.)
- 5.16. Goffin, J.L. "On the Convergence Rates of Subgradient Optimization Methods." *Math. Progr.* 13, 3 (1977): 329-47.
- 5.17. Kelley, J.E., Jr. "The Cutting Plane Method for Solving Convex Programs." *J. SIAM* 8, 4 (1960): 703-12.
- 5.18. Newman, D.J. "Location of the Maximum on Unimodal Surfaces." *J. ACM* 12, 3 (1965): 395-98.
- 5.19. Balinski, M.L., and Wolfe, P., eds. *Nondifferentiable Optimization. Math. Progr. Study* 3. Amsterdam: North-Holland, 1975.
- 5.20. Lemarechal, C., and Mifflin, R., eds. *Nonsmooth Optimization. Proc. IIASA Workshop*, 28 March-8 April 1977. Oxford: Pergamon Press, 1978.



English translation exists

- 5.21.** Polyak, B.T. "Sharp Minimum." A Talk Given at the IIASA Workshop on Generalized Lagrangians and Their Applications, 17-22 December 1979, IIASA, Luxenburg, Austria.
- 5.22.** Ashchepkov, L.T., Belov, et al. *Metody Resheniya Zadach Matematicheskogo Programmirovaniya i Optimal'nogo Upravleniya* (Methods for Solving Problems of Mathematical Programming and Optimal Control). Novosibirsk: Nauka, 1984.
- 5.23.** Bland, R.G., Goldfarb, D., and Todd, M.J. "The Ellipsoid Method: A Survey." *Oper. Res.* 29, 6 (1981): 1039-91.
- 5.24.** Ecker, J.G., and Kupferschmid M. "An Ellipsoid Algorithm for Non-linear Programming." *Math. Progr.* 27, 1 (1983): 83-106.
- 5.25.** Gershovich, V.I., and Shor, N.Z. "Method of Ellipsoids, Its Generalizations and Applications." *Cybernetics*, vol. 18, no. 5 (1982): 602-17. (English transl.)
- 5.26.** Goffin, J.-L. "Convergence Rates of the Ellipsoid Method on General Convex Functions." *Math. Oper. Res.* 8, 1 (1983): 135-50.
- 5.27.** Goldfarb, D., and Todd, M.J. "Modifications and Implementation of the Ellipsoid Algorithms for Linear Programming." *Math. Progr.* 23, 1 (1982): 1-19.
- 5.28.** Konig, H., and Pallaschke, D. "On Khachian's Algorithm and Minimal Ellipsoid." *Numer. Math.* 36, 2 (1981): 211-23.
- 5.29.** Lemarechal, C. "Nondifferentiable Optimization." In *Nonlinear Optimization: Theory and Algorithms*, ed. L.C.W. Dixon, E. Spedicato, and G.P. Szego, 149-99. Boston: Birkhauser, 1980.
- 5.30.** Nesterov, Yu.E. "Minimization Methods for Convex and Quasiconvex Functions" (in Russian). *Ekonomika i Matemat. Metody* 20, 3 (1984): 519-31.
- 5.31.** Nurminskij, E.A., ed. *Progress in Nondifferentiable Optimization*. IIASA, Luxenburg, Austria, 1982.
- 5.32.** Rockafellar, R.T. *The Theory of Subgradients and its Application to Problems of Optimization: Convex and Nonconvex Functions*. Berlin: Heldermann, 1981.
- 5.33.** Topkis, D.M. "A Cutting-plane Algorithm With Linear and Geometric Rates of Convergence." *J. Optim. Theory Appl.* 36, 1 (1982): 1-22.
- 5.34.** Yannitskij, B., and Levin, L.A. "An Old Linear Programming Algorithm Runs in Polynomial Time." *Proc. 23d Ann. IEEE/FOCS Symp.*, 327-28. Chicago, Silver Springs, Maryland, 1982.

## Chapter 6

(see also 0.2 0.8, 0.11, 0.12, 0.16, 1.3, 1.4, 1.8)

- 6.1.** A. Albert. *Regression and the Moore-Penrose pseudoinverse*. New York: Academic Press, 1972.
- 6.2.** Bakushinskij (Bakusinskii), A.B., and Polyak (Poljak), B.T. "On the Solution of Variational Inequalities." *Soviet Math. Dokl.*, vol. 15, no. 6 (1974): 1705-10. (English transl.)

- 6.3.** Batishchev, D.I. *Poiskovye Metody Optimal'nogo Proektirovaniya* (Search Methods of Optimal Planning). Moscow: Sovetskoe Radio, 1975.
- 6.4.** Bulatov, V.P. *Metody Pogruzheniya v Zadachakh Optimizatsii* (Methods of Imbedding in Optimization Problems). Novosibirsk: Nauka, 1977.
- 6.5.** Gel'fand, I.M., and Tsetlin, M.L. "Principle of Nonlocal Search in Systems of Automatic Optimization" (in Russian). *Doklady Akademii Nauk SSSR*, Ser. Kibernetika i Teoriya Regulirovaniya, vol. 137, no. 2 (1961): 295-98.
- 6.6.** Emelin, I.V., and Krasnosel'skij (Krasnosel'skii), M.A. "On the Theory of Ill-posed Problems." *Soviet Math. Dokl.*, vol. 20, no. 1 (1979): 105-109. (English transl.)
- 6.7.** Eremin, I.I., and Mazurov, V.D. *Nonstationary Processes of Mathematical Programming* (in Russian). Moscow: Nauka, 1979.
- 6.8.** Lavrent'ev, M.M. *Some Ill-posed Problems of Mathematical Physics*. Berlin New York: Springer-Verlag, 1967.
- 6.9.** Levitin, E.S., and Polyak (Poljak), B.T. "Convergence of Minimizing Sequences in Conditional Extremum Problems." *Soviet Math. Dokl.*, vol. 7, no. 3 (1966): 764-67. (English transl.)
- 6.10.** Morozov, V.A. "On Regularization of Certain Classes of Extremal Problems" (in Russian). In *Vychislitel'nye Metody i Programmirovaniye* 12 (1969): 24-37. Moscow: Izdatel'stvo MGU.
- 6.11.** I.B. Motskus. *Mnogoekstremal'nye Zadachi v Proektirovaniii* (Multi-extremal Problems in Planning). Moscow: Nauka, 1967.
- 6.12.** B.T. Polyak. "Iterative Methods of Solving Ill-posed Variational Problems" (in Russian). In *Vychislitel'nye Metody i Programmirovaniye* 12 (1969): 38-52. Moscow: Izdatel'stvo MGU.
- 6.13.** Rastrigin, L.A. *Statisticheskie Metody Poiska* (Statistical Search Methods). Moscow: Nauka, 1968.
- 6.14.** Sobol', I.M. *Mnogomernye Kvadraturnye Formuly i Funktsii Haara* (Multidimensional Quadrature Formulas and Haar's Functions). Moscow: Nauka, 1969.
- 6.15.** Strongin, R.G. *Chislennye Metody v Mnogoekstremal'nykh Zadachakh* (Numerical Methods in Multiextremal Problems). Moscow: Nauka, 1978.
- 6.16.** Sukharev, A.G. *Optimal'nyj Poisk Ekstremuma* (Optimal Extremum Search). Moscow: Izdatel'stvo MGU, 1975.
- 6.17.** Fedorova, I.E. "Search for a Global Optimum in Multiextremal Problems" (in Russian). In *Teoriya Optimal'nykh Reshenij*, vyp. 4, pp. 93-101. Vilnius, USSR, Institut Mekhaniki i Kibernetiki, 1978.
- 6.18.** Khejsin, V.E. "Iterative Procedures of Minimization Under Conditions of Drift of the Extremum" (in Russian). *Avtomatika i Telemekhanika* 11 (1976): 91-101.
- 6.19.** Martinet, B. "Regularization of Variational Inequalities by Successive Approximations" (in French). *RAIRO*, vol. 4, no. R3 (1970): 154-59.



English translation exists

- 6.20.** Rockafellar, R.T. "Monotone Operators and the Proximal Point Algorithm." *SIAM J. Contr. Optim.*, vol. 14, no. 5 (1976): 877-98.
- 6.21.** Dixon, L.C.W., and Szego, G.P., eds. *Towards Global Optimization*. Amsterdam New York: North-Holland, 1975.
- 6.22.** Zilinskas, A. "On Statistical Models for Multimodal Optimization." *Math. Operations Statist.*, ser. Statist., vol. 9, no. 2 (1978): 255-66.
- 6.23.** Jongen, H. Th., Jonver, P., and Twilt, F. *Nonlinear Optimization in R<sup>n</sup>*. Frankfurt am Main Berne New York: Peter Lang, 1983.
- 6.24.** Nemirovskij, A.S., and Polyak, B.T. "Iterative Methods for Solving Linear Ill-posed Problems under Exact Information, I, II" (in Russian). *Tekhnicheskaya Kibernetika* 2 (1983): 13-25; 3 (1983): 18-25.
- 6.25.** Nesterov, Yu.E. "A Method of Solving a Convex Programming Problem with Convergence Rate  $O(1/k^2)$ ." *Soviet Math. Dokl.*, vol. 7, no. 3 (1983): 372-76. (English transl.)
- 6.26.** Polyak, B.T. "Iterative Algorithms for Singular Minimization Problems." In *Nonlinear Programming*, edited by O.L. Mangasarian, R.R. Meyer, and S.M. Robinson, 147-68. New York: Academic Press, 1981.

### Chapter 7

(see also 0.2, 0.4, 0.8, 0.9, 0.13-0.15, 3.4, 3.9, 4.1, 4.2, 4.5, 5.1, 5.8-5.12, 5.15, 5.21)

- 7.1.** Gupal, A.M., and Bazhenov, L.G. "A Stochastic Linearization." *Cybernetics* 3 (1972): 482-84. (English transl.)
- 7.2.** Bertsekas, D.P. "On the Goldstein-Levitin-Polyak Gradient Projection Method." *IEEE Trans. Autom. Control*, vol. 21, no. 2 (1976): 174-84.
- 7.3.** Cannon, M.D., and Cullum, C.D. "A Tight Upper Bound on the Rate of Convergence of the Frank-Wolfe Algorithm." *SIAM J. Control*, vol. 6, no. 4 (1968): 509-16.
- 7.4.** Dunn, J.C. "Rates of Convergence for Conditional Gradient Algorithms Near Singular and Nonsingular Extremals." *SIAM J. Contr. Optim.*, vol. 17, no. 2 (1979): 187-211.
- 7.5.** Frank, M., and Wolfe, P. "Algorithm for Quadratic Programming." *Naval Res. Log. Quart.*, vol. 3, nos. 1-2 (1956): 95-110.
- 7.6.** Goldstein, A.A. "Convex Programming in Hilbert Space." *Bull. Amer. Math. Soc.*, vol. 70, no. 5 (1964): 709-10.
- 7.7.** Bertsekas, D.P. "Enlarging the Region of Convergence of Newton's Method for Constrained Optimization." *J. Optim. Theory Appl.*, vol. 36, no. 2 (1982): 221-52.
- 7.8.** Dunn, J.C. "Global and Asymptotic Convergence Rate Estimates for a Class of Projected Gradient Processes." *SIAM J. Contr. Optim.*, vol. 19, no. 3 (1981): 368-400.



English translation exists

**Chapter 8**

(see also 0.13, 0.15, 0.20-0.23)

- 8.1.** Antipin, A.S. "A Gradient-type Method for Finding the Saddle Point of the Augmented Lagrangian" (in Russian). *Ekonomika i Matemat. Metody*, vol. 13, no. 3 (1977): 560-65.
- 8.2.** Volin, Yu.M., and Ostrovskij, G.M. "The Penalty Function Method and Necessary Optimality Conditions" (in Russian). In *Upravlyayemye Sistemy*, vyp. 9 (1971): 43-51. Novosibirsk: SO AN SSSR.
- 8.3.** Levitin, E.S. "Differentiability with Respect to a Parameter of the Optimal Value in Parametric Problems of Mathematical Programming." *Cybernetics*, vol. 12, no. 1 (1976): 46-64. (English transl.)
- 8.4.** Lyusternik, L.A., and Sobolev, V.I. *Elementy Funktsional'nogo Analiza* (Fundamentals of Functional Analysis). 2nd ed. Moscow: Nauka, 1965.
- 8.5.** Pervozvanskij, A.A., and Gajtsgori, V.G. *Decomposition, Aggregation, and Approximate Optimization* (in Russian). Moscow: Nauka, 1979.
- 8.6.** Polyak, B.T. "Iterative Methods Using Lagrange Multipliers for Solving Extremal Problems with Constraints of the Equation Type." *U.S.S.R. Comput. Maths. Math. Phys.*, vol. 10, no. 5 (1970): 42-52. (English transl.)
- 8.7.** ——. "The Convergence Rate of the Penalty Function Method." *U.S.S.R. Comput. Maths. Math. Phys.*, vol. 11, no. 1 (1971): 1-12. (English transl.)
- 8.8.** ——. "Methods for Solving Constrained Extremum Problems in the Presence of Random Noise." *U.S.S.R. Comput. Maths. Math. Phys.*, vol. 19, no. 1 (1979): 72-81. (English transl.)
- 8.9.** Polyak, B.T., and Tret'yakov, N.V. "The Method of Penalty Estimates for Conditional Extremum Problems." *U.S.S.R. Comput. Maths. Math. Phys.*, vol. 13, no. 1 (1973): 42-58. (English transl.)
- 8.10.** Pshenichnyj, B.N. "Algorithms for the General Mathematical Programming Problem" (in Russian). *Kibernetika* 5 (1970): 120-25.
- 8.11.** Bertsekas, D.P. "Multiplier Methods: A Survey." *Automatica* 12 (1976): 133-45.
- 8.12.** ——. *Constrained Optimization and Lagrange Multiplier Methods*. New York London: Academic Press, 1982.
- 8.13.** Haarhoff, P.C., and Buys, J.D. "A New Method for the Optimization of a Nonlinear Function Subject to Nonlinear Constraints." *Comput. J.*, vol. 13, no. 2 (1970): 178-84.
- 8.14.** Hestenes, M.R. "Multiplier and Gradient methods." *J. Optim. Theory Appl.*, vol. 4, no. 5 (1969): 303-30.
- 8.15.** Powell, M.J.D. "A Method for Nonlinear Constraints in Minimization Problems." In *Optimization*, edited by R. Fletcher, 283-98. London: Academic Press, 1969.
- 8.16.** Robinson, S.M. "A Quadratically Convergent Algorithm for General Nonlinear Programming Problems." *Math. Progr.*, vol. 3, no. 2 (1972): 145-56.

- 8.17.** P. Wolfe. *Methods for Nonlinear Constraints*. In *Nonlinear Programming*, edited by J. Abadie, 120-31. Amsterdam: North-Holland, 1967.
- 8.18.** Boggs, P.T., and Tolle, J.W. "Augmented Lagrangians Which are Quadratic in the Multiplier." *J. Optim. Theory Appl.*, vol. 31, no. 1 (1980): 17-26.
- 8.19.** Di Pillo, G., and Grippo, L. "A New Class of Augmented Lagrangians in Nonlinear Programming." *SIAM J. Contr. Optim.*, vol. 17, no. 1 (1979): 618-28.
- 8.20.** Pshenichnyj, B.N. *The Linearization Method* (in Russian). Moscow: Nauka, 1983.

## Chapter 9

(see also 0.1-0.3, 0.5-0.11, 0.13-0.15, 0.17-0.23)

- 9.1.** Alekseev, V.M., Tikhomirov, V.M., and Fomin, S.V. *Optimal'noe Upravlenie* (Optimal Control). Moscow: Nauka, 1979.
- 9.2.** Antipin, A.S. *Methods of Nonlinear Programming Based on the Direct and Dual Augmentation of the Lagrangian* (in Russian). Moscow: VNIISI, 1979. Preprint.
- 9.3.** Girsanov, I.V. *Lectures on Mathematical Theory of Extremum Problems*. Berlin Heidelberg New York: Springer-Verlag, 1972. (English transl.)
- 9.4.** Gol'stejn, E.G. *Teoriya Dvojstvennosti v Matematicheskem Programmirovani i Ee Prilozheniya* (Duality Theory in Mathematical Programming and Its Applications). Moscow: Nauka, 1971. German transl.: *Dualitätstheorie in der Nichtlinearen Optimierung und Ihrer Anwendung*. Berlin: Akademie-Verlag, 1975.
- 9.5.** ——. "On the Convergence of the Gradient Method for Finding the Saddle Points of Augmented Lagrangians" (in Russian). *Ekonom. i Matemat. Metody*, vol. 13, no. 2 (1977): 322-29.
- 9.6.** Gol'stejn, E.G., and Tret'yakov, N.V. "Augmented Lagrangians" (in Russian). *Ekonom. i Matemat. Metody*, vol. 10, no. 3 (1974): 568-91.
- 9.7.** Dubovitskii, A.Ya., and Milyutin, A.A. "Extremum Problems in the Presence of Restrictions." *U.S.S.R. Comput. Maths. Math. Phys.*, vol. 5, no. 3 (1965): 1-80. (English transl.)
- 9.8.** Eremin, I.I. "On the Penalty Method in Convex Programming" (in Russian). *Kibernetika* 4 (1967): 63-67.
- 9.9.** Eremin, I.I., and Astaf'ev, N.N. *Vvedenie v Teoriyu Linejnogo i Vypuklogoprogrammirovaniya* (Introduction to the Theory of Linear and Convex Programming). Moscow: Nauka, 1976.
- 9.10.** Zhukhovitskij, S.I., and Avdeeva, L.I. *Linejnoe i Vypukloe Programmirovaniye* (Linear and Convex Programming). Moscow: Nauka, 1967.
- 9.11.** Zuhovitskij (Zuhovickii), S.I., Polyak (Poljak), R.A., and Primak, M.E. "An Algorithm for Solution of the Convex Programming Problem." *Soviet Math. Dokl.*, vol. 4, no. 1-6 (1963): 1754-57. (English transl.)

- 9.12.** Ioffe, A.D., and Tikhomirov, V.M. *Theory of Extremal Problems*. Amsterdam New York: North-Holland, 1979. (English transl.)
- 9.13.** Lebedev, V.Yu. "Convergence of the Weighted Functional Method in Convex Programming Problems." *U.S.S.R. Comput. Maths. Math. Phys.*, vol. 17, no. 3 (1977): 198-202. (English transl.)
- 9.14.** Majstrovskij, G.D. "On Gradient Methods for Finding Saddle Points" (in Russian). *Ekonomika i Matemat. Metody*, vol. 12, no. 5 (1976): 917-29.
- 9.15.** ——. "On the Rate of Convergence of the Gradient Method for an Augmented Lagrangian" (in Russian). *Ekonomika i Matemat. Metody*, vol. 15, no. 2 (1979): 380-86.
- 9.16.** Polyak, R.A. "An Algorithm for Simultaneous Solution of the Primal and Dual Problems of Convex Programming" (in Russian). In *Ekonomicheskaya Kibernetika i Issledovanie Operatsij*, no. 3, pp. 53-64. Trudy Seminara, Kiev, Institut Kibernetiki AN SSSR, 1966.
- 9.17.** Pshenichnyj, B.N. "Dual Method in Extremal Problems, I, II." (in Russian). *Kibernetika* 3 (1965): 89-95; 4 (1965): 64-69.
- 9.18.** Tret'yakov, N.V. "The Penalty Function Method for Convex Programming Problems" (in Russian). *Ekonomika i Matemat. Metody*, vol. 9, no. 3 (1973): 526-40.
- 9.19.** Elster, K.-H., and Grossmann, C. "Solution of Nonlinear Optimization Problems Using Penalty and Barrier Functions" (in Russian). In *Primenenie Issledovaniya Operatsij v Ekonomike*, 95-161. Moscow: Economika, 1977.
- 9.20.** Bertsekas, D.P. "On the Method of Multipliers for Convex Programming." *IEEE Trans. Autom. Control*, vol. 20 (1975): 385-88.
- 9.21.** Garcia-Palomares, U.M., and Mangasarian, O.L. "Superlinearly Convergent Quasi-Newton Algorithms for Nonlinearly Constrained Optimization Problems." *Math. Progr.*, vol. 11, no. 1 (1976): 1-13.
- 9.22.** Grossmann, C., and Kaplan, A.A. *Penalty Method and Modified Lagrangian in Nonlinear Optimization*. Leipzig, DDR: BSB B.G. Teubner Verlagsgesellschaft, 1979.
- 9.23.** Han, S.P. "Superlinearly Convergent Variable Metric Algorithm for General Nonlinear Programming Problem." *Math. Progr.*, vol. 11, no. 3 (1976): 263-82.
- 9.24.** Kowalik, J., Osborne, M.R., and Ryan, D.M. "A New Method for Constrained Optimization Problems." *Oper. Res.*, vol. 7, no. 6 (1969): 973-83.
- 9.25.** Kuhn, H.W. "Nonlinear Programming: a Historical View." In *Nonlinear Programming. SIAM-AMS Proc.*, vol. 9, 1-26. Providence, R.I.: American Math. Society, 1976.
- 9.26.** Mangasarian, O.L. *Nonlinear Programming*. New York: McGraw-Hill, 1969.
- 9.27.** Morrison, D.D. "Optimization by Least Squares." *SIAM J. Numer. Anal.*, vol. 5, no. 1 (1968): 83-88.
- 9.28.** Neustadt, L.W. *Optimization. A Theory of Necessary Conditions*. Princeton, New Jersey: Princeton University Press, 1976.

- 9.29.** Pietrzykowski, T. "An Exact Potential Method for Constrained Maxima." *SIAM J. Numer. Anal.*, vol. 16, no. 2 (1969): 299-304.
- 9.30.** Rockafellar, R.T. "The Multiplier Method of Hestenes and Powell Applied to Convex Programming." *J. Optim. Theory Appl.*, vol. 12 (1973): 555-62.
- 9.31.** Rosen, J.B. "The Gradient Projection Method for Nonlinear Programming. I." *SIAM J.*, vol. 8, no. 1 (1960): 180-217.
- 9.32.** Wierzbicki, A.P. "A Penalty Function Shifting Method in Constrained Static Optimization and Its Convergence Properties." *Archiw. Autom. i Telemech.*, vol. 16, no. 4 (1971): 395-416.

## Chapter 10

(see also 0.7, 0.19, 0.22, 2.1, 5.15, 5.21, 6.2, 8.1, 8.12, 9.9, 9.10, 9.15)

- 10.1.** Bulavskij, V.A., Zvyagina, R.A., and Yakovleva, M.A. *Chislennye Metody Linejnogo Programmirovaniya* (Numerical Methods of Linear Programming). Moscow: Nauka, 1977.
- 10.2.** Gass, Saul I. *Linear Programming: Methods and Applications*. New York: McGraw-Hill Book Co., 1958.
- 10.3.** Gabasov, R., and Kirillova, F.M. *Methods of Linear Programming* (in Russian). Minsk, USSR: Izdatel'stvo Belorus. Gos. Universiteta, 1977 (vols. 1, 2); 1980 (vol. 3).
- 10.4.** Gol'stejn, E.G., and Yudin, D.B. *Novye Napravleniya v Linejnem Programmirovaniyu* (New Trends in Linear Programming). Moscow: Sovetskoe Radio, 1966.
- 10.5.** Dantzig, G.B. *Linear Programming and Extensions*. Princeton, New Jersey: Princeton University Press, 1963.
- 10.6.** Kantorovich, L.V. *Matematicheskie Metody v Organizatsii i Planirovaniyu Proizvodstva* (Mathematical Methods in Organizing and Planning Production). Leningrad: Izdatel'stvo Leningradskogo Gos. Universiteta, 1939.
- 10.7.** Karpelevich, F.I., and Sadovskij, L.E. *Elements of Linear Algebra and Linear Programming* (in Russian). Moscow: Fizmatgiz, 1963.
- 10.8.** Korpelevich, G.M. "The Extragradient Method for Finding Saddle Points and Related Problems" (in Russian). *Ekonomika i Matemat. Metody*, vol. 12, no. 4 (1976): 747-56.
- 10.9.** Künzi, Hans Paul, and Krelle, Wilhelm. *Nonlinear Programming*. Waltham, Mass.: Blaisdell, 1966.
- 10.10.** Kuhn, H.W., and Tucker, A.W., eds. *Linear Inequalities and Related Systems*. Annals of Math. Studies, vol. 38. Princeton, New Jersey: Princeton University Press, 1956.
- 10.11.** Nit, I.V. *Linear Programming* (includes also some nonlinear problems) (in Russian). Moscow: Izdatel'stvo Moskovsk. Gos. Universiteta, 1978.

- 10.12.** Polyak, B.T., and Tret'yakov, N.V. "On an Iterative Method of Linear Programming and Its Economic Interpretation" (in Russian). *Ekonomika i Matemat. Metody*, vol. 8, no. 5 (1972): 740-51.
- 10.13.** Razumikhin, B.S. *Fizicheskie Modeli i Metody Teorii Ravnovesiya v Programmirovani i Ekonomike* (Physical Models and Equilibrium Theory Methods in Mathematical Programming and Economics). Moscow: Nauka, 1975.
- 10.14.** Romanovskij, I.V. *Algoritmy Resheniya Ekstremal'nykh Zadach* (Algorithms for Solving Extremal Problems). Moscow: Nauka, 1977.
- 10.15.** Khachiyan (Hacijan), L.G. "A Polynomial Algorithm in Linear Programming." *Soviet Math. Dokl.*, vol. 20, no. 1 (1979): 191-94. (English transl.)
- 10.16.** Yudin, D.B., and Gol'stejn, E.G. *Zadachi i Metody Linejnogo Programmirovaniya* (Problems and Methods of Linear Programming). Moscow: Sovetskoe Radio, 1961.
- 10.17.** ——. *Linejnoe Programmirovanie. Teoriya i Konechnye Metody* (Linear Programming. Theory and Finite Methods). Moscow: Fizmatgiz, 1963.
- 10.18.** Klee, V., and Minty, G.J. "How Good is the Simplex Algorithm?" In *Inequalities*, III, 159-75, edited by O. Shisha. New York London: Academic Press, 1972.
- 10.19.** Adler, I., Megiddo, N., and Todd, M.J. "New Results on the Average Behavior of Simplex Algorithms." *Bull. AMS*, vol. 11, no. 2 (1984): 378-82.
- 10.20.** Borgwardt, K.H. "The Average Number of Steps Required by the Simplex Method is Polynomial." *Z. Oper. Res.*, Ser. A-B, vol. 26, no. 5 (1982): 157-77.
- 10.21.** Karmarkar, N. "A New Polynomial Time Algorithm for Linear Programming." *Combinatorica*, vol. 4, no. 4 (1984): 373-95.
- 10.22.** Mangasarian, O.L. "Iterative Solution of Linear Programs." *SIAM J. Numer. Anal.*, vol. 18, no. 4 (1981): 606-14.
- 10.23.** Megiddo, N. "Linear Programming in Linear Time When the Dimension is Fixed." *J. AMS*, vol. 31, no. 1 (1984): 114-27.
- 10.24.** Smale, Steve. "On the Average Number of Steps of the Simplex Method of Linear Programming." *Math. Progr.*, vol. 27, no. 3 (1983): 241-62.
- 10.25.** Vershik, A.M., and Sporyshev, P.V. "An Estimate of the Average Number of Steps in the Simplex Method, and Problems in Asymptotic Integral Geometry." *Soviet Math. Dokl.*, vol. 28, no. 1 (1983): 195-99. (English transl.)

## Chapter 11

(see also 0.1, 0.19, 4.2-4.4, 4.9, 4.11, 6.3, 6.11, 6.13, 10.5, 10.6, 10.13)

- 11.1.** Anderson, T. *The Statistical Analysis of Time Series*. New York: John Wiley Publ. Co., 1971.
- 11.2.** Aris, Rutherford. *Discrete Dynamic Programming; An Introduction to the Optimization of Staged Processes*. New York: Blaisdell, 1964.

- 11.3.** Bard, Jonathan. *Nonlinear Parameter Estimation*. New York: Academic Press, 1974.
- 11.4.** Bellman, Richard E. *Dynamic Programming*. Princeton, New Jersey: Princeton University Press, 1957.
- 11.5.** Vapnik, Vladimir N. *Estimation of Dependences Based on Empirical Data*. New York Berlin Heidelberg: Springer-Verlag, 1982. (English transl.)
- 11.6.** Mikhalevich, V.C., ed. *Vychislitel'nye Metody Vybora Optimal'nykh Proektivnykh Reshenij* (Computational Methods of Choosing Optimal Design Solutions). Kiev: Naukova Dumka, 1977.
- 11.7.** Gemintern, V.I., and Kagan, B.M. *Metody Optimal'nogo Proektirovaniya* (Methods of Optimal Design). Moscow: Energiya, 1980.
- 11.8.** Germejer, Yu.B. *Vvedenie v Teoriyu Issledovaniya Operatsij* (An Introduction to Operations Research Theory). Moscow: Nauka, 1971.
- 11.9.** Gurin, L.S., Dymarskij, Ya.S., and Merkulov, A.D. *Zadachi i Metody Optimal'nogo Raspredeleniya Resursov* (Problems and Methods Optimal Allocation of Resources). Moscow: Sovetskoe Radio, 1968.
- 11.10.** Duffin, Richard J., and Zener, Clarence. *Geometric Programming*. New York: John Wiley and Sons Publ. Co., 1967.
- 11.11.** Dennis, Jack Bonnell. *Mathematical Programming and Electrical Networks*. Cambridge Technology Press of the Massachusetts Institute of Technology, 1959.
- 11.12.** Ermol'ev, Yu.M., and Yastremskij, A.I. *Sokhasticheskie Modeli i Metody v Ekonomicheskem Planirovaniy* (Stochastic Models and Methods in Economic Planning). Moscow: Nauka, 1979.
- 11.13.** Ershov, A.A. "Stable Methods of Parameter Estimation: A Survey" (in Russian). *Avtomatika i Telemekhanika* 8 (1978): 66-100.
- 11.14.** Kazakevich, V.V., and Rodov, A.B. *Systems of Automatic Optimization* (in Russian). Moscow: Energiya, 1977.
- 11.15.** Kantorovich, L.V. *Ekonomicheskij Raschet Nailuchshego Ispol'zovaniya Resursov* (Economic Optimization of the Use of Resources). Moscow: AN SSSR, 1960.
- 11.16.** Loran, P.J. *Approximation et Optimisation*. Paris: Hermann, 1972.
- 11.17.** Lasdon, L.S. *Optimization Theory of Large-scale Systems*. London: Macmillan, 1971.
- 11.18.** Mudrov, V.I., and Kushko, V.L. *Metody Obrabotki Izmerenij* (Methods of Processing Measurements). Moscow: Sovetskoe Radio, 1976.
- 11.19.** Nalimov, V.V., and Chernova, N.A. *Statisticheskie Metody Planirovaniya Ekstremal'nykh Eksperimentov* (Statistical Methods of Design of Extremal Experiments). Moscow: Nauka, 1965.
- 11.20.** Ostrovskij, G.M., and Volin, V.M. *Methods of Optimizing Chemical Reactors* (in Russian). Moscow: Khimiya, 1967.
- 11.21.** Pervozvanskij, A.A. *Matematicheskie Modeli v Upravlenii Proizvodstvom* (Mathematical Models in Production Control). Moscow: Nauka, 1975.



English translation exists

- 11.22.** Polyak, B.T., and Skokov, V.A. "Solutions of Minimum Problems for Sums of Squares" (in Russian). *Ekonomika i Matemat. Metody*, vol. 14, no. 6 (1978): 1173-80.
- 11.23.** Polyak, B.T., and Tsyplkin, Ya.Z. "Adaptive Algorithms for Estimation (Convergence, Optimality, Stability)" (in Russian). *Avtomatika i Tele-Mekhanika* 3 (1979): 71-84.
- 11.24.** Propoj, A.I. *Elementy Teorii Optimal'nykh Diskretnykh Protsessov* (Fundamentals of the Optimal Discrete Process Theory). Moscow: Nauka, 1973.
- 11.25.** Tabak, D., and Kuo, B.C. *Optimal Control by Mathematical Programming*. Englewood Cliffs, New Jersey: Prentice-Hall, 1971.
- 11.26.** Ul'm, S.Yu. *Metody Dekompozitsii dlya Resheniya Zadach Optimizatsii* (Decomposition Methods of Solving Optimization Problems). Tallin, USSR: Valgus, 1979.
- 11.27.** Fedorov, V.V. *Theory of Optimal Experiment* (in Russian). Moscow: Nauka, 1971.
- 11.28.** ——. *Chislennye Metody Maksimina* (Numerical Maximin Methods). Moscow: Nauka, 1979.
- 11.29.** Himmelblau, David Mautner. *Process Analysis by Statistical Methods*. New York: John Wiley & Sons Publ. Co., 1970.
- 11.30.** Tsirlin, A.M. *Variatsionnye Metody Vybora Optimal'nykh Parametrov Apparatov Khimicheskoy Tekhnologii* (Variational methods of Choosing Optimal Parameters in Chemical Engineering). Moscow: Mashino-stroenie, 1978.
- 11.31.** Chernous'ko, F.L., and Banichuk, N.V. *Variatsionnye Zadachi Mekhaniki i Upravleniya. Chislennye Metody*. (Variational Problems of Mechanics and Control). Moscow: Nauka, 1973.
- 11.32.** Bertsekas, D.P. "Approximation Procedures Based on the Method of Multipliers." *J. Optim. Theory Appl.*, vol. 23, no. 4 (1977): 487-510.
- 11.33.** Bracken, Jerome, and McCormick, G.P. *Selected Applications of Nonlinear Programming*. New York: John Wiley & Sons Publ. Co., 1968.
- 11.34.** Lasdon, L.S., and Waren, A.D. "Survey of Nonlinear Programming Applications." *Oper. Res.*, vol. 28, no. 5 (1980): 1029-73.
- 11.35.** Athanari, T.S., and Dodge, Y. *Mathematical Programming in Statistics*. New York: John Wiley & Sons Publ. Co., 1981.
- 11.36.** Bradley, S.P., Hax, A.C., and Magnanti, T.L. *Applied Mathematical Programming*. Reading, Mass.: Addison-Wesley Publ. Co., 1977.
- 11.37.** Daniel, G., and Wood, F.S. *Fitting Equations to Data*. New York: John Wiley & Sons Publ. Co., 1980.
- 11.38.** Dneprovskij (Dneprovskii), I.E. "On a Method of Solving Problems of Best Discrete Approximation." *U.S.S.R. Comput. Mats. Math. Phys.*, vol. 24, no. 5 (1984): 124-30. (English transl.)
- 11.39.** Heine, R., and Petry, K. "Bibliographie zur geometrischen Optimierung." *Math. Optim. Stat., Ser. Optim.*, vol. 14, no. 3 (1983): 467-80.

- 11.40.** Huber, Peter J. *Robust Statistics*. New York: John Wiley & Sons Publ. Co., 1981.
- 11.41.** Kashyap, R.L., and Rao, A.R. *Dynamic Stochastic Models From Empirical Data*. New York: Academic Press, 1976.
- 11.42.** Zanakis, S.H., and Rustagi, Jagdish S. *Optimization in Statistics*. Amsterdam: North-Holland, 1982.
- 11.43.** Polyak (Poljak), B.T. "On the Bertsekas Method for Minimization of Composite Functions." *Proc. International Symposium on Systems Optimization and Analysis*, Rocquencourt, December 11-13, 1978, edited by A. Bensoussan and J.L. Lions. Lecture Notes in Control and Information Sci., vol. 14, 179-86. Berlin Heidelberg New York: Springer-Verlag, 1979.

## Chapter 12

(see also 0.21, 10.14)

- 12.1.** Belov, E.N. "Algorithms and Programs for Solving Problems of Quadratic and Linear Programming" (in Russian). *Ekonomika i Matemat. Metody*, vol. 16, no. 1 (1980): 198-201.
- 12.2.** Belov, E.N., Polyak, B.T., and Skokov, V.A. "A Complex of Optimization Programs" (in Russian). *Ekonomika i Matemat. Metody*, vol. 14, no. 4 (1978): 792-96.
- 12.3.** Nesterov, Yu.E., and Skokov, V.A. "On the Problem of Testing Unconstrained Minimization Algorithms." In *Chislennye Metody Matemat. Programmirovaniya*, 77-91. Moscow: TsEMI, 1980.
- 12.4.** Malkov, U.Kh. "Survey of Programs for Solving the General Problem of Linear Programming" (in Russian). *Ekonomika i Matemat. Metody*, vol. 5, no. 4 (1969): 594-97.
- 12.5.** Skokov, V.A. "Computational Experiments in Solving Nonlinear Programming Problems." In *Matemat. Metody Resheniya Ekonomicheskikh Zadach. Sbornik* 7, 51-69. Moscow: Nauka, 1977.
- 12.6.** Moiseev, N.N., ed. *Sovremennoe Sostoyanie Teorii Issledovaniya Operatsij* (The Status of Operations Research Theory). Moscow: Nauka, 1979.
- 12.7.** Wilkinson, J.H., and Reinsch, C. *Linear Algebra. Handbook for Automatic Computation*. New York Berlin Heidelberg: Springer-Verlag, 1971.
- 12.8.** Balinski, M.L., and Hellerman, Eli, eds. *Computational Practice in Mathematical Programming. Math. Progr. Study*, vol. 4. Amsterdam: North-Holland, December 1975.
- 12.9.** Colville, A.R. "A Comparative Study on Nonlinear Programming Codes." *Proc. Princeton Sympos. Math. Programming*, edited by H.W. Kuhn. Princeton, New Jersey: Princeton University Press, 1970.
- 12.10.** Crowder, H.P., Dembo, R.S., and Mulvey, J.M. "Reporting Computational Experiments in Mathematical Programming." *Math. Progr.*, vol. 15, no. 3 (1978): 316-29.

- 12.11.** Dembo, R.S. "A Set of Geometric Programming Test Problems and Their Solutions." *Math. Progr.*, vol. 10, no. 2 (1976): 192-213.
- 12.12.** Kuester, J.L., and Mize, J.H. *Optimization Techniques with FORTRAN*. New York: McGraw-Hill Book Co., 1973.
- 12.13.** Künzi, H.P., Tzsachach, H.G., and Zender, C.A. *Numerical Methods of Mathematical Optimization with ALGOL and FORTRAN Programs*. New York: Academic Press, 1971.
- 12.14.** Land, A.H., and Powell, S. *FORTRAN Codes for Mathematical Programming: Linear, Quadratic and Discrete*. London New York: John Wiley & Sons Publ. Co., 1973.
- 12.15.** Orchard-Hays, W. *Advanced Linear Programming Computing Techniques*. New York: McGraw-Hill Book Co., 1968.
- 12.16.** Waren, A.D., Lasdon, L.S. "The Status of Nonlinear Programming." *Oper. Res.*, vol. 27, no. 3 (1979): 431-56.
- 12.17.** Hock, Willie, and Schittkowski, Klaus. *Test Examples for Nonlinear Programming Codes*. Lecture Notes in Econom. and Math. Systems, vol. 187. Berlin Heidelberg New York, 1978.
- 12.18.** Kim, K.V., Nesterov, Yu.E., et al. "An Effective Algorithm for Computing Derivatives and Extremal Problems" (in Russian). *Ekonomika i Matemat. Metody*, vol. 20, no. 2 (1984): 309-318.
- 12.19.** Moré, J.J., Garbow, B.S., and Hillstrom, K.E. "Testing Unconstrained Optimization Software." *ACM Trans. Math. Software*, vol. 7, no. 1 (March 1981): 17-41.
- 12.20.** Schittkowski, Klaus. *Nonlinear Programming Codes*. Lect. Notes in Econom. and Math. Systems, vol. 183. Berlin Heidelberg New York: Springer-Verlag, 1980.

## INDEX

Basis 318

Cone 256

Convergence

    Convergence almost surely (with probability 1) 48

    Convergence in probability 48

    Convergence in the mean square 48

    Linear convergence 40

Convex hull 121

Derivative 2

    Directional derivative 4

Descent

    Direction of steepest descent 138

    Local steepest descent 21

Differentiability

    Fréchet differentiability 4

    Gâteaux differentiability 4

Direction of recession 308

Epigraph 128

Feasible solution

Function

    Affine function 9

    Augmented Lagrange function 232

    Concave function 9

    Convex function 8

    Strictly convex function 9

    Strongly convex function 9

    Differentiable function 3

    Differentiable at a point 5

    Differentiable on a set 3

    Gâteaux differentiable function 4

    Homogeneous function 84

    Lyapunov function 51

    Powell's function 386

    Regularization function 173

    Rosenbrock's function 385

    Scalar function 2

    Shor's function 391

    Twice differentiable scalar function 6

- Hessian 6
- Gradient 2
- Growth Condition 51
- Inequality
  - Chebyshev inequality 48
  - Cramer-Rao inequality 116
  - Jensen's inequality 8
  - Kantorovich inequality 62
  - Kolmogorov's inequality 48
- Iterative regularization 176
- Lagrange multipliers 225
- Lemma
  - Caratheodory's lemma 121
  - Farkas' lemma 258
  - Gladyshev lemma 49
  - Moreau-Rockafellar' lemma 256
  - Robbins-Siegmund lemma 50
- Mapping 56
  - Contraction mapping 56
  - Contraction mapping principle 56
- Martingale 48
- Matrix 5
  - Matrix of second derivatives 6
  - Fisher information matrix 116
  - Hessian matrix 6
  - Jacobian matrix 5
  - Condition number of a matrix 19
- Method
  - Artificial basis method 320
  - Augmented Lagrangian method 241, 328
  - Barrier method 288
  - Broyden method 77
  - Broyden-Fletcher-Shanno method 77
  - Conjugate-gradient method 68, 105
  - Cutting-plane method 146
  - Davidon-Fletcher-Powell method 77
  - Direct methods 87
  - Difference analog of the gradient method 88
  - Dual methods 301
  - Dual simplex method 322

- Ellipsoid method 155  
Exterior penalty method 288  
Fibonacci's method 162  
Gauss-Seidel method 90  
Heavy-ball method 65  
Interior penalty method 288  
Keifer-Wolfowitz method 106  
Levenberg-Marquardt method 63  
Monte-Carlo method 96  
Penalty function method 288, 327  
Penalty shifting method 289  
Proximal (prox-) method 175  
Random search method 114  
Reduced gradient method 301  
Regularization method 173, 181, 326  
Rosen's gradient projection method 281  
Search methods 87  
Secant method 81  
Simplex method 320  
Simplicial method 91  
Steepest descent method 60  
Zero-order methods 87  
Method of barycentric coordinates 92  
Method of Chebyshev centers 149  
Method of coordinatewise descent 88  
Method of difference approximation of the gradient ~~method~~ 106  
Method of feasible directions 280  
Method of random coordinatewise descent 89  
Method of random search 89  
Method of simultaneous solution of the primal and dual problems 287  
Methods of variable metric 80  
Methods without derivatives 87  
Minimum  
    Global minimum 11, 200  
    Local minimum 11  
Monotonicity condition  
    Strict monotonicity condition 10  
    Strong monotonicity condition 10

- Necessary first-order minimum condition 224  
Necessary second-order minimum condition 230  
Noise, types of 97  
    Absolute deterministic noise 97  
    Absolute random noise 97  
    Relative deterministic noise 97  
    Relative random noise 97  
Normal solution of a problem 174
- Overrelaxation 90
- Point  
    Admissible point 224  
    Minimum point 11, 224  
    Globally stable minimum point 17  
    Local minimum point 200  
    Locally stable minimum point 16  
    Locally unique minimum point 15  
    Nonsingular minimum point 15  
    Regular minimum point 225  
    Sharp minimum point 136, 205  
    Weakly stable minimum point 17  
    Stationary point 12  
    Nonsingular stationary point 185  
    Index of a stationary point 185
- Polar of a cone 257
- Posynomial 358
- Problem  
    Dual problem 265  
    Generalized stable problem 204  
    Geometric programming problem 358  
    Primal problem 265  
    Quadratic programming problem 334
- Proximal operator 175
- Pseudogradient 51
- Pseudoinverse 174
- Regularization parameter 173
- Ridge index 18
- Rule for differentiation of composite functions 5
- Rule of Lagrange multipliers 225

Stability margin 18  
Subgradient 127  
Submartingale 48  
Sufficient first-order minimum condition 201  
Supermartingale 48  
Supporting hyperplane 124

Theorem

Duality theorem 265, 310, 311  
Fermat theorem 11  
Hahn-Banach theorems 122  
Helly's theorem 373  
Implicit function theorem 57  
Karush-John's theorem 271  
Kuhn-Tucker's theorem 260  
Nemirovskij-Yudin's theorem 153  
Rademacher theorem 128  
Separation theorem 122  
Steiner's theorem 374  
Weierstrass theorem 15, 204  
Theorem on supporting hyperplane 124

Variation 4

Vector

$A$ -orthogonal vector 70  
Conjugate vector 70

Vertex 306

Nonsingular vertex 317