Introduction to Optimization

Book · July 2020

CITATIONS
322

READS
9,214

1 author:

Boris T. Polyak
Institute of Control Sciences
287 PUBLICATIONS **16,019** CITATIONS

SEE PROFILE

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

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

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

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
Massachussetts 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. Tsypkin, 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;

$\{x_1,...,x_n\}$ are the components of the vector $x \in \mathbf{R}^n$;

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

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

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 $y \in \mathbf{R}^n$, $x_i \geq y_i$, $i = 1,...,n$;

$\mathbf{R}^n_+$ 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 $x \in \mathbf{R}^n$: $(x_+)_i = \max\{0,x_i\}$, $i = 1,...,n$;

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

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

$\nabla f(x), f'(x)$ 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);

$L'_x(x,y)$, $L''_{xx}(x,y)$: 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 ε-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);

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: $\rho(x,Q) = \inf\limits_{y \in Q} \|x - y\|$

$o(h(x))$: if $g:\mathbf{R}^n \to \mathbf{R}^m$, $h:\mathbf{R}^n \to \mathbf{R}^s$ and $\|g(x)\|/\|h(x)\| \to 0$ as $\|x\| \to 0$, then $g(x) = o(h(x))$;

$O(h(x))$: if $g:\mathbf{R}^n \to \mathbf{R}^m$, $h:\mathbf{R}^n \to \mathbf{R}^s$ and there are $\varepsilon > 0$, α such that $\|g(x)\| \le \alpha\|h(x)\|$ for $\|x\| \le \varepsilon$, then $g(x) = O(h(x))$;

$o(u_k)$: if the sequences $u_k \in \mathbf{R}^n$, $v_k \in \mathbf{R}^m$, $k = 1,2,...$, are such that $\|v_k\|/\|u_k\| \to 0$ as $k \to \infty$, then $v_k = o(u_k)$;

$O(u_k)$: if for sequences $u_k \in \mathbf{R}^n$, $v_k \in \mathbf{R}^m$, $k = 1,2,...$, there are $\alpha > 0$, such that $\|v_k\| \le \alpha\|u_k\|$ for $k \ge k_0$, then $v_k = O(u_k)$;

$\mathbf{E}\xi$ is the mathematical expectation of the random variable ξ;

$\mathbf{E}(\xi|x)$ is the conditional mathematical expectation of the random variable ξ depending on x for a fixed value of x;

$\forall$ is the universal quantifier: $\forall x \in Q$ means "for all $x \in Q$";

$\square$ 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; α, β, ... 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, ..., x^k, ...; 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 ($f: \mathbf{R}^n \to \mathbf{R}^1$) is said to be *differentiable at a point* x if we can find a vector $a \in \mathbf{R}^n$ such that for all $y \in \mathbf{R}^n$,

$$f(x + y) = f(x) + (a,y) + o(y) . \tag{1}$$

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

$$f(x + y) = f(x) + (\nabla f(x), y) + o(y) . \tag{2}$$

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 $(\partial f(x)/\partial x_1, \ldots, \partial f(x)/\partial x_n)$. 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 $|(Ay,y)| \le \|A\| \, \|y\|^2$. Hence $(Ay,y)/2 = o(y)$. Thus, $f(x)$ is differentiable at any point x and

$$\nabla f(x) = Ax - b . \tag{3}$$

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 $x + \tau y$, $0 \le \tau \le 1$). We consider the one-variable function $\phi(\tau) = f(x + \tau y)$ and compute its derivative for $0 \le \tau \le 1$:

$$\begin{aligned} \frac{\phi(\tau+\Delta\tau) - \phi(\tau)}{\Delta\tau} &= \frac{f(x+(\tau+\Delta\tau)y) - f(x+\tau y)}{\Delta\tau} \\ &= \frac{(\nabla f(x+\tau y), \Delta\tau y) + o(\Delta\tau y)}{\Delta\tau} , \end{aligned}$$

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

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

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

The quantity

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

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), $a = (\partial f/\partial x_1, ..., \partial f/\partial x_n)$. It follows from formula (4) that if $f(x)$ is differentiable at x, then it is also Gâteaux differentiable, with

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

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

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

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:

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

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):

$$f(x + y) = f(x) + (\nabla f(x + \theta y), y) , \tag{9}$$

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: $\mathbf{R}^n \to \mathbf{R}^m$ is said to be *differentiable at a point* x if we can find an $m \times n$-matrix A such that for all $y \in \mathbf{R}^n$,

$$g(x + y) = g(x) + Ay + o(y) . \tag{10}$$

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

$$g(x + y) = g(x) + g'(x)y + o(y) , \tag{11}$$

i.e., a function differentiable at x admits at x a first-order linear approximation. Obviously, for a differentiable vector function $g(x) = (g_1(x), ..., g_m(x))$ 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: $\mathbf{R}^n \to \mathbf{R}^m$ be differentiable at x, and h: $\mathbf{R}^m \to \mathbf{R}^s$ be differentiable at $g(x)$. Then the *rule for differentiatiation of composite functions* (or *chain rule*) is valid:

$$[h(g(x))]' = h'(g(x))\, g'(x) , \tag{12}$$

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 θ, $0 \leq \theta \leq 1$, such that

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

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

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

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

$$\|g(x+y) - g(x)\| \le L\|y\| , \tag{14}$$

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

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

then

$$\|g(x+y) - g(x) - g'(x)y\| \le L\|y\|^2/2 . \tag{15}$$

As in the scalar case, a function g: $\mathbf{R}^n \to \mathbf{R}^m$ 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 $y \in \mathbf{R}^n$,

$$f(x+y) = f(x) + (\nabla f(x), y) + (Hy, y)/2 + o(\|y\|^2)$$

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

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

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

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

In particular, if

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

we have

$$|f(x+y) - f(x) - (\nabla f(x), y)| \le (L/2)\|y\|^2 , \tag{19}$$

whereas if

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

then

$$|f(x+y) - f(x) - (\nabla f(x), y) - (\tfrac{1}{2})(\nabla^2 f(x)y, y)| \le (L/6)\|y\|^3 . \tag{20}$$

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

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

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

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

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 $f''(x) = (f'(x))^T$, 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

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda) f(y) \tag{22}$$

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.

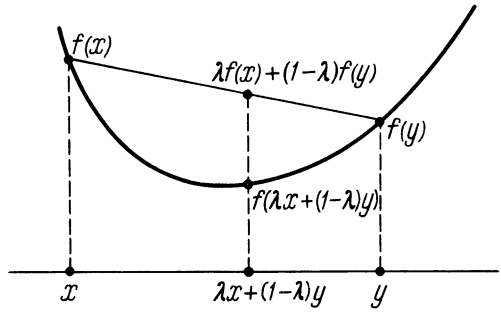


Fig. 1 A convex function.

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

$$f(\lambda_1 x^1 + \cdots + \lambda_k x^k) \leq \lambda_1 f(x^1) + \cdots + \lambda_k f(x^k) \,. \quad \square \tag{23}$$

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, $i = 1,...,m$, then $f(x) = \Sigma_{i=1}^{m} \gamma_i f_i(x)$, $\gamma_i \geq 0$, and $f(x) = \max\limits_{1 \leq i \leq m} f_i(x)$ 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 $x \neq y$, $0 < \lambda < 1$,

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1 - \lambda)f(y) , \tag{24}$$

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

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

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 $\psi(\tau)$ 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 ($\psi'(\tau_1) - \psi'(\tau_2) \geq \ell(\tau_1 - \tau_2)$, $\tau_1 > \tau_1$). □

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

$$f(x + y) \geq f(x) + (\nabla f(x), y) , \tag{26}$$

strict convexity to the inequality

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

and strong convexity to the inequality

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

for any $x, y \in \mathbf{R}^n$. □

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).

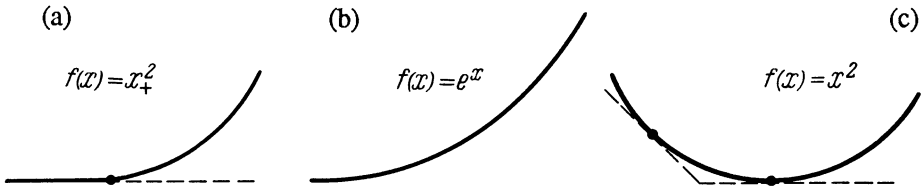


Fig. 2 Types of convexity: (a) a convex function;
(b) a strictly convex function;
(c) a strongly convex function.

From (26) we obtain the useful inequality

$$(\nabla f(x) - \nabla f(y), x-y) \geq 0 , \tag{29}$$

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:

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

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

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

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

$$\nabla^2 f(x) \geq 0 , \tag{32}$$

and strong convexity is equivalent to the condition

$$\nabla^2 f(x) \geq \ell I \tag{33}$$

for all x. If

$$\nabla^2 f(x) > 0 \tag{34}$$

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 ℓ). 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

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

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

$$\|\nabla f(x)\| \geq \ell\|x - x^*\| . \tag{37}$$

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 ε-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

$$\nabla f(x^*) = 0 . \tag{1}$$

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. □

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.

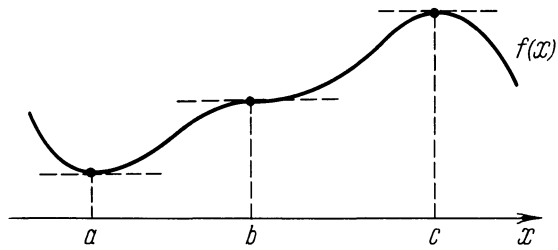


Fig. 3 Stationary points: a is a minimum point, b is an inflection point; c is a maximum point.

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 $x \in \mathbf{R}^n$. □

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

$$\nabla^2 f(x^*) \geq 0 . \tag{2}$$

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

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

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

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

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

$$\nabla^2 f(x^*) > 0 . \tag{3}$$

Then x^* is a local minimum point.

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

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

$$\geq f(x^*) + \tau^2 \ell/2 + o(\tau^2) ,$$

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 τ_0 such that for $0 \leq \tau \leq \tau_0$ we have $\tau^2\ell/2 \geq o(\tau^2)$, i.e., $f(x^* + \tau y) \geq f(x^*)$. □

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 $Q_\alpha = \{x: f(x) \le \alpha\}$ for some α be nonempty and bounded. Then there exists a global minimum point of $f(x)$ on $\mathbf{R}^n$.

PROOF. Let

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

Then $x^k \in Q_\alpha$ for sufficiently large k. The set Q_α 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.,

$$x^* = \operatorname*{arg\,min}_{x \in \mathbf{R}^n} f(x) . \quad \square$$

The assumption of the boundedness of Q_α 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 $f(x) = (Ax,x) - (b,x)$, $A \ge 0$, and let $f(x)$ be bounded below (e.g., $f(x) \ge 0$). Prove that $f(x)$ attains its minimum on $\mathbf{R}^n$, although the conditions of Theorem 1 do not generally hold (the set Q_α 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

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

$$\geq \ell \|x-x^*\| + o(\|x-x^*\|) > 0$$

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. □

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. □

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 $f(x^k) \to f(x^*)$, $\|x^k - x^*\| \leq \delta$ imply $x^k \to x^*$.

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, $\|x^k - x^*\| \leq \delta$, $f(x^k) \to f(x^*)$. By the compactness of a unit sphere in $\mathbf{R}^n$, one can take a convergent subsequence $x^{k_i} \to \bar{x}$, $\|\bar{x} - x^*\| \leq \delta$. 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. □

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_ε in a neighborhood of x^* and $x_\varepsilon \to x^*$ as $\varepsilon \to 0$. □

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_ε of the function $f(x) + \varepsilon g(x)$ in a neighborhood of x^*, and

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

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 $Q_\alpha = \{x: f(x) \le \alpha\}$ is nonempty and bounded for some $\alpha > f(x^*)$, then x^* is globally stable. The requirement for the Q_α 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 $x^k \to \infty$ 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

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

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

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

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

Thus, the number ℓ characterizes the "stability margin" of a minimum point. However, ℓ 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

$$\mu = \overline{\lim_{\delta \to 0}} \left(\sup_{x \in L_\delta} \|x - x^*\|^2 \Big/ \inf_{x \in L_\delta} \|x - x^*\|^2 \right), \tag{4}$$

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

the ridge index of a minimum point x^*. In other words, μ 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 μ 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 μ 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 μ is relevant to many problems involving unconstrained minimization and can serve as a measure of the complexity of the problem.

For a quadratic function

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

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 ℓ_1 is the normalized eigenvector corresponding to the smallest eigenvalue λ_1 of the matrix A and the factor γ_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 $x_n = x^* + \gamma_n \ell_n$, ℓ_n being the eigenvector corresponding to the largest eigenvalue λ_n, $\gamma_n = (2\delta/\lambda_n)^{1/2}$ (Fig. 4). Thus the ratio

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

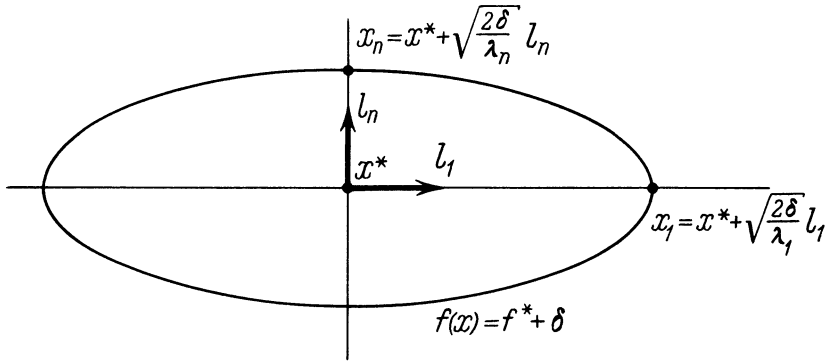


Fig. 4 Condition number of a quadratic function.

√ to the smallest

does not really depend on δ and

$$\mu = \frac{\lambda_n}{\lambda_1}. \tag{6}$$

Note that in Linear Algebra the ratio of the largest eigenvalue 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

$$\mu = \frac{L}{\ell}, \tag{7}$$

where L is the largest eigenvalue and ℓ 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 $Q_\alpha = \{x: f(x) \le \alpha\}$ 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 ϕ: $\mathbf{R}^1 \to \mathbf{R}^1$ 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

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

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 ε_k and solve the auxiliary problem

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

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

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

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

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

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 $\gamma_k \equiv \gamma$:

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

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

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

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| , \tag{6}$$

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

$$f(x) \geq f^* > -\infty , \tag{7}$$

and let γ satisfy the condition

$$0 < \gamma < 2/L . \tag{8}$$

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

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

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

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

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

Summing the inequalities

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

from 0 to s over k, we obtain

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

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

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

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

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 $\|\nabla f(x) - \nabla f(0)\|/\|x - 0\| = (\alpha + 1)\|x\|^{\alpha - 1} \to \infty$ as $\|x\| \to 0$. 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 $\|\nabla f(x) - \nabla f(0)\|/\|x - 0\| = (2 + \alpha)\|x\|^{\alpha} \to \infty$ as $\|x\| \to \infty$. 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 $\|x^{k+1}\| > \|x^k\|$, $k = 0,1,\ldots$.

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 γ, violating condition (8), as is seen from $f(x) = Lx^2/2$, $x \in \mathbf{R}^1$. Indeed, if $\gamma \geq 2/L$, then in method (5) for this function one has $f(x^{k+1}) \geq f(x^k)$, $k = 0,1,\ldots$, 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 $\|x^k\| \to \infty$ for any $x^0 \neq 0$.

If we require that $\{x: f(x) \leq f(x^0)\}$ 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 γ_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

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

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

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

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

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

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., $A \ge 0$. Moreover, $\|A\| \le L$ since $\|\nabla^2 f(x)\| \le L$ for all x by a Lipschitz condition on the gradient. Hence

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

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

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

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 ℓ. Then for $0 < \gamma < 2/L$ method (5) converges to a unique global minimum point x^* with the rate of geometric progression:

$$\|x^k - x^*\| \le cq^k , \quad 0 \le q < 1 . \tag{12}$$

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

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

We use Lemma 3:

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

yielding

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

$$= q_1(f(x^k) - f(x^*)) ,$$

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

Since $0 < \gamma < 2/L$, then $0 < q_1 < 1$, and therefore $f(x^k) \to f(x^*)$. 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

$$\ell I \leq \nabla^2 f(x) \leq LI , \qquad \ell > 0 , \tag{13}$$

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

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

The quantity q is minimal and equal to

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

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

$$\|x^{k+1} - x^*\| = \|x^k - x^* - \gamma\nabla f(x^k)\| = \|(I - \gamma A_k)(x^k - x^*)\|$$

$$\leq \|I - \gamma A_k\|\, \|x^k - x^*\| .$$

For every symmetric matrix A we have $\|I - A\| = \max\{|1 - \lambda_1|, |1 - \lambda_n|\}$, where λ_1 and λ_n are respectively the smallest and the largest eigenvalues of A. Hence $\|x^{k+1} - x^*\| \le q\|x^k - x^*\|$, $q = \max\{|1 - \gamma\ell|, |1 - \gamma L|\}$. Since $0 < \gamma < 2/L$, $0 < \ell \le L$, then $|1 - \gamma\ell| < 1$, $|1 - \gamma L| < 1$, i.e., $q < 1$. Minimizing q over γ, we obtain (15). □

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

$$f(x) = (Ax, x)/2 - (b, x), \qquad A > 0,\ 0 < \ell = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n = L,$$

where the λ_i are the eigenvalues of A. Take an arbitrary $0 < \gamma < 2/L$. Assume that $|1 - \gamma\ell| \ge |1 - \gamma L|$. Take $x^0 = x^* + e^1$, where e^1 is the normalized eigenvector corresponding to λ_1. Then

$$x^k - x^* = (I - \gamma A)^k (x^0 - x^*) = (1 - \gamma\lambda_1)^k e^1,$$

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

Similarly, if $|1 - \gamma L| \ge |1 - \gamma\ell|$, take $x^0 = x^* + e^n$, where e^n is the normalized eigenvector corresponding to λ_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

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

cannot be improved even if γ 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 $q = \max\{|1 - \gamma\ell|, |1 - \gamma L|\} \le q^*$ only for $\gamma = \gamma^*$, 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 $\|x^0 - x^*\| \leq \varepsilon$,

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

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

The quantity q is minimal and equal to

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

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 γ does the method converge? For which does it diverge?

ANSWERS: (a) No convergence for any $\gamma > 0$ and $x^0 \neq 0$, with $|x^k| \to [(1/2)(1 + \alpha)\gamma]^{1/(1-\alpha)}$ 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 $|x^k| \to \infty$ 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 $|x^k| \to \infty$ for $\gamma > 2$. (d) $|x^k| \to \infty$ 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,

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

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 minumum point of $f_k(x)$ as the new approximation:

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathbf{R}^n} f_k(x) .$$

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 ℓ, 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. □

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. □

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.

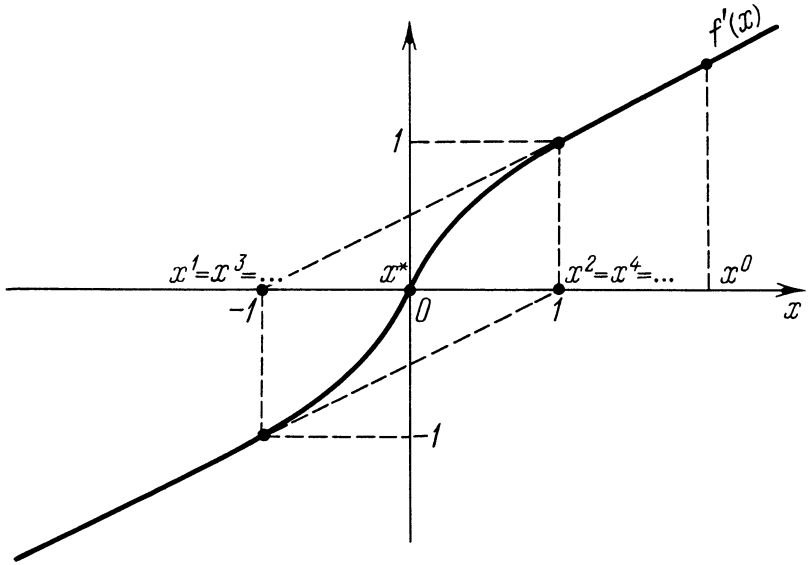


Fig. 5 Divergence of Newton's method.

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:

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

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 kth iteration, yielding

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

THEOREM 3. Let equation (7) have a solution x^*, let a function $g: \mathbf{R}^n \to \mathbf{R}^n$ 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 $\|x^0 - x^k\| \le \varepsilon$ 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. □

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 $\nabla f(x^k) \to 0$ (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, $\|x^k - x^*\| \le 2q^{2^k}$ (since $\ell \not\geq L$). Hence for $q = 0.5$, we have $\|x^k - x^*\| \le 2^{-2^k+1}$, so that $\|x^5 - x^*\| < 10^{-9}$, whereas for $q = 0.1$, we have $\|x^k - x^*\| \le 2 \cdot 10^{-2^k}$, 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 γ, by virtue of Theorem 3 of Section 1.4, converges geometrically with ratio $q = (L-\ell)/(L+\ell)$, and we saw that this estimate was exact for the qua-

dratic 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 μ 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 ℓ, 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 $\gamma = 2/(L + \ell)$

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

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 $x^k \to x^*$? 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, γ and q in those theorems are indeed usually unknown and therefore a constructive choice of γ in the gradient method or explicit estimates of the rate of convergence are impossible. There are however more complicated ways of choosing γ_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 $\lambda_1, \ldots, \lambda_n$ be its eigenvalues. By the spectral radius of A we mean the quantity

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

The norm

$$\|A\| = \max_{\|x\|=1} \|Ax\| \tag{2}$$

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, $\rho(A) \le \|A\|$ and generally $\rho(A) \ne \|A\|$. For example, for the matrix $A = \left(\begin{smallmatrix} 0 & 0 \\ 1 & 0 \end{smallmatrix}\right)$, 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

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

which implies the following lemma.

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

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

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

$$A^T U A = U - C \tag{4}$$

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

PROOF. Since $\|A^k\| \le cq^k$, $q < 1$ (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, $U \ge 0$ for $C \ge 0$,

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

if $C \ge 0$. □

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

$$\operatorname{Re} \lambda_i < 0, \quad i = 1,\ldots,n. \tag{5}$$

LEMMA 3. For $\lim_{t\to\infty} e^{At} = 0$ 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 $\|e^{At}\| \le c(\varepsilon) e^{(\gamma+\varepsilon)t}$ for all $t \ge 0$, $\gamma = \max_i \operatorname{Re} \lambda_i$.

Indeed, the eigenvalues of $B = e^A$ are e^{λ_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). □

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

$$AU + UA^T = -C \tag{6}$$

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 $C > 0$ ($C \geq 0$). □

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

LEMMA 5. Let A be stable,

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

Then $\rho(B) < 1$.

Indeed, if λ_i are the eigenvalues of A and μ_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., $\rho(B) < 1$. □

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 $A^k x^0 \to 0$ as $k \to \infty$.
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 kth iteration.

Consider an iterative process of the form

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

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 $\rho = \sqrt{(g'(x^*))} < 1$. 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 $k \geq 0$

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

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

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

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

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

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

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

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

Then, if

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

where U is the solution of the matrix equation

$$UA + A^T U = I , \tag{12}$$

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

$$\begin{aligned} \|x^k\|^2 &\le \|x^0\|^2 \, \|U^{-1}\| \, \|U\| q^k , \\ q &= 1 - (\tfrac{1}{2})\gamma\|U\|^{-1} + \gamma L + (\tfrac{1}{2})\gamma^2(\|A\| + L)^2 . \end{aligned} \tag{13}$$

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

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

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

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

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

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

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

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. 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 $S = \{x: \|x - x^*\| \le \|x^0 - x^*\|\}$, let $g'(x)$ satisfy a Lipschitz condition on S, and let $g'(x^*) = 0$. Then, if

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

then

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

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

$$\|x^1 - x^*\| = \|g(x^0) - g(x^*) - g'(x^*)(x^0 - x^*)\|$$

$$\le (L/2)\|x^0 - x^*\|^2 \le q\|x^0 - x^*\|.$$

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

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

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

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

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 $V(x^k) - V(x^{k+1}) \to 0$. 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

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

holds at the kth step of the process. Hence the conclusion that $u_k \to 0$ and the estimate of the rate of convergence of u_k. The behavior of sequences of the form (1) for certain "typical" functions ϕ_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

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

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

$$u_{k+1} \le c u_k^2 , \quad c > 0 , \tag{3}$$

yielding $u_k \le c^{-1}(cu_0)^{2^k}$ and if $cu_0 < 1$ then $u_k \to 0$.

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

We start with linear inequalities of the form

$$u_{k+1} \le q_k u_k + \alpha_k , \quad q_k \ge 0 , \tag{4}$$

implying

$$u_k \le q_{k-1} q_{k-2} \cdots q_0 u_0 + q_{k-1} \cdots q_1 \alpha_0 + \cdots + q_{k-1}\alpha_{k-2} + \alpha_{k-1} . \tag{5}$$

Now we consider some special cases.

LEMMA 1. Let

$$u_{k+1} \le q u_k + \alpha , \quad 0 \le q < 1 , \ \alpha > 0 . \tag{6}$$

Then

$$u_k \le \alpha/(1-q) + (u_0 - \alpha/(1-q))q^k . \tag{7}$$

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

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

LEMMA 2. Let $u_k \ge 0$ and let

$$u_{k+1} \le (1 + \alpha_k)u_k + \beta_k , \quad \alpha_k \ge 0 , \ \beta_k \ge 0 ,$$
$$\sum_{k=0}^{\infty} \alpha_k < \infty , \quad \sum_{k=0}^{\infty} \beta_k < \infty . \tag{8}$$

Then $u_k \to u \ge 0$.

The proof is the same as that of the more general Lemma 9 below. □

LEMMA 3. Let

$$u_{k+1} \le q_k u_k + \alpha_k, \quad 0 \le q_k < 1, \quad \alpha_k \ge 0,$$
$$\sum_{k=0}^{\infty} (1 - q_k) = \infty, \quad \alpha_k/(1 - q_k) \to 0. \tag{9}$$

Then $\overline{\lim_{k\to\infty}} u_k \le 0$. In particular, if $u_k \not\to 0$, then $u_k \to 0$. □

COROLLARY. If in (9) $q_k \equiv q < 1$, $\alpha_k \to 0$, $u_k \ge 0$, then $u_k \to 0$. □

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

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

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

Then

$$u_k \le d(c-p)^{-1} k^{-p} + o(k^{-p}) \quad \text{for } c > p, \tag{11}$$
$$u_k = O(k^{-c} \log k) \quad \text{for } p = c, \tag{12}$$
$$u_k = O(k^{-c}) \quad \text{for } p > c. \tag{13}$$

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

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

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

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

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

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

$$v_{k+1} = u_{k+1}(k+1)^c \le \left[\left(1-\frac{c}{k}\right)u_k + \frac{d}{k^{p+1}}\right] k^c \left(1 + \frac{c}{k} + \frac{c^2}{2k^2} + o\left(\frac{1}{k^2}\right)\right)$$

$$= \left(1 - \frac{c^2}{2k^2} + o\left(\frac{1}{k^2}\right)\right)v_k + \frac{d}{k^{p-c+1}}\left(1 + O\left(\frac{1}{k}\right)\right) \le v_k + \frac{d'}{k^{p-c+1}}$$

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

LEMMA 5 (Chung). Let $u_k \ge 0$

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

Then

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

We proceed to investigate recurrence inequalities defined by nonlinear relations.

LEMMA 6. Let $u_k > 0$ and let

$$u_{k+1} \le u_k - \alpha_k u_k^{1+p}, \quad \alpha_k \ge 0, \ p > 0. \tag{15}$$

Then

$$u_k \le u_0\left[1 + pu_0^p \sum_{i=0}^{k-1} \alpha_i\right]^{-1/p}. \tag{16}$$

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

$$u_k \le u_0/(1 + \alpha k u_0). \tag{17}$$

PROOF. We have

$$0 < u_{k+1} \le u_k(1 - \alpha_k u_k^p), \quad 1 - \alpha_k u_k^p > 0,$$

$$u_{k+1}^{-p} \ge u_k^{-p}(1 - \alpha_k u_k^p)^{-p} \ge u_k^{-p}(1 + p\alpha_k u_k^p) = u_k^{-p} + p\alpha_k.$$

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

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

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

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

$\phi(u) > 0$ 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 $u_k \to 0$.

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 $u_k \to \bar{u} \geq 0$. Since $\phi(u)$ is monotone, $\phi(u_k) \geq \phi(\bar{u})$, $u_{k+1} \leq u_k - (1-\theta)\gamma_k\phi(\bar{u})$. Summing the inequalities and taking into account the assumption $\Sigma\gamma_k = \infty$ we get $\phi(\bar{u}) = 0$, i.e., $\bar{u} = 0$.

(b) I_1 is infinite. Then for $k \in I_1$, $\phi(u_k) \leq \varepsilon_k + \delta_k u_k$, $\varepsilon_k = \beta_k/(\theta\gamma_k) \to 0$, $\delta_k = \alpha_k/(\theta\gamma_k) \to 0$. If $\alpha_k \equiv 0$, then $\delta_k \equiv 0$, $\phi(u_k) < \varepsilon_k \to 0$, hence $u_k \to 0$ for $k \in I_1$, $k \to \infty$. If $\alpha_k \not\equiv 0$ 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 $u^*(\varepsilon, \delta) \to 0$ for $\varepsilon, \delta \to 0$. 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 $u_k \to 0$ for $k \in I_1$, $k \to \infty$. Thus $u_k \to 0$ for $k \in I_1$, $k \to \infty$ regardless $\alpha_k \equiv 0$ or $\alpha_k \not\equiv 0$. If $k \in I_1$, $k+1 \in I_2$, then $u_{k+1} \leq (1+\alpha_k)u_k + \beta_k \to 0$ for $k \in I_1$, $k \to \infty$. Finally, if $k \in I_1$, $k+j \in I_2$, $j = 1, \ldots, s$, then $u_{k+j} \leq u_{k+j-1} \leq \cdots \leq u_{k+1}$. Hence $u_k \to 0$, $k \to \infty$.

Thus $u_k \to 0$ for both cases. □

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 $v^1, \ldots, v^k, \ldots$ be a sequence of n-dimensional random vectors. We shall not specify the probability space $(\Omega, \mathcal{F}, \mathbf{P})$ on which these variables are defined (i.e., we do not write $v^1(\omega), \ldots, v^k(\omega), \omega \in \Omega$, Ω being the space of elementary events, $\mathcal{F}$ being the σ-algebra of measurable sets defined on it, $\mathbf{P}$ 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 $\mathbf{P}(\lim_{k\to\infty} v^k = v) = 1$ (here and in the sequel, $\mathbf{P}(A)$ denotes the probability of the event A), and we indicate this by $v^k \to v$ a.s.;

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

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

The theory of semimartingales is the basic tool for studying convergence of random variables. A sequence of scalar random variables $v_0, \ldots, v_k, \ldots$ is called a *supermartingale* if $\mathbf{E}(v_{k+1}|v_0,\ldots,v_k) \le v_k$, $\mathbf{E}v_0 < \infty$, where $\mathbf{E}(v_{k+1}|v_0,\ldots,v_k)$ is the conditional mathematical expectation of v_{k+1} for the given $v_0, \ldots, v_k$. 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 $v_0, \ldots, v_k, \ldots$ be a supermartingale, where $v_k \ge 0$ for all k. Then there is a random variable $v \ge 0$ such that $v_k \to v$ a.s. □

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

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

$$\mathbf{P}(\exists k: v_k \ge \varepsilon) \le \varepsilon^{-1}\mathbf{E}v_0 \,. \quad \square \tag{18}$$

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

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

$$\mathbf{E}(v_{k+1} \mid v_0,\ldots,v_k) \leq (1 + \alpha_k)v_k + \beta_k ,$$

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

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

PROOF. Introduce

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

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

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

Here

$$\mathbf{E}(u_{k+1} \mid u_0,\ldots,u_k)$$

$$= \prod_{i=k+1}^{\infty} (1 + \alpha_i)\, \mathbf{E}(v_{k+1} \mid v_0,\ldots,v_k) + \sum_{i=k+1}^{\infty} \beta_i \prod_{j=i+1}^{\infty} (1 + \alpha_j)$$

$$\leq \prod_{i=k}^{\infty} (1 + \alpha_i)v_k + \sum_{i=k}^{\infty} \beta_i \prod_{j=i+1}^{\infty} (1 + \alpha_j) = u_k ,$$

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

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

LEMMA 10. Let $v_0, \ldots, v_k$ be a sequence of random variables, $v_k \geq 0$, $\mathbf{E}v_0 < \infty$ and let

$$\mathbf{E}(v_{k+1} \mid v_0, \ldots, v_k) \leq (1 + \alpha_k)v_k + \beta_k , \tag{20}$$

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

$$\frac{\beta_k}{\alpha_k} \to 0 . \tag{21}$$

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

$$\mathbf{P}(v_j \le \varepsilon \text{ for all } j \ge k) \ge 1 - \varepsilon^{-1}\left(\mathbf{E}v_k + \sum_{i=k}^{\infty} \beta_i\right). \tag{22}$$

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

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

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

In the preceding lemmas the quantities α_k, β_k were deterministic. Consider the case when they are random (and perhaps dependent).

LEMMA 11 (Robbins-Siegmund). Let v_k, u_k, α_k, β_k be nonnegative random variables and let

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

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

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

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

where $v \ge 0$ is some random variable. □

2.2.3 The Main Theorems

Consider an iterative process of the form

$$x^{k+1} = x^k - \gamma_k s^k, \tag{23}$$

where k is the number of the iteration, x^k, s^k are vectors in $\mathbf{R}^n$, $\gamma_k \ge 0$ 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}, ... 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 $V(x)$:

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

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:

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

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

e) The initial approximation satisfies the condition

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

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

f) The step size is such that

$$\gamma_k \geq 0 , \quad \sum_{k=0}^{\infty} \gamma_k = \infty , \quad \overline{\lim_{k \to \infty}} \gamma_k < \frac{2}{L\tau} . \tag{27}$$

Let us state the basic convergence theorems. Under conditions a-f, it is generally impossible to assert that $V(x^k) \to 0$ for process (23) in any probabilistic sense. For example, if $s^k \equiv 0$, then all the conditions hold, but $x^k \equiv x^0$. 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 $\Sigma_{k=0}^{\infty} \gamma_k^2 < \infty$. Then for any x^0 in algorithm (23) one has

$$V(x^k) \to V \quad \text{a.s.,} \quad \underline{\lim_{k \to \infty}} (\nabla V(x^k), \mathbf{E}(s^k \mid x^k)) = 0 . \tag{28}$$

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:

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

By conditions c and f we have

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

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

$$\mathbf{E}V(x^{k+1}) \leq \mathbf{E}V(x^k) - \gamma_k(1 - (\tfrac{1}{2})L\tau\gamma_k)u_k + L\gamma_k^2\sigma^2/2 ,$$

$$u_k = \mathbf{E}(\nabla V(x^k), \mathbf{E}(s^k \mid x^k)) .$$

For sufficiently large k, by condition f, we have

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

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

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

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

c')
$$(\nabla V(x^k), \mathbf{E}(s^k \mid x^k)) \geq \ell V(x^k) , \qquad \ell > 0 .$$

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

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

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

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

The required result follows from Lemma 10 and condition f. □

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 $\gamma_k \to 0$. Then in algorithm (23) one has

$$\mathbf{E}V(x^k) \to 0 . \quad (33)$$

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

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

Since

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

for sufficiently large k, then

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

By Lemma 3, $\mathbf{E}V(x^k) \to 0$. □

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

$$\mathbf{E}V(x^k) \leq \mathbf{E}V(x^0)q^k + \frac{L\gamma\sigma^2}{\ell(2-L\tau\gamma)}(1-q^k) , \quad (35)$$

$$q = 1 - \ell\gamma(1-(\tfrac{1}{2})L\tau\gamma) .$$

This result follows from (34) and Lemma 1. □

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

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

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

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

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

This result can easily be derived from (34) and Lemma 4. □

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:

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

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

Under certain conditions imposed on ℓ_k, β_k, σ_k, τ_k and μ_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 $V(x^k) \to 0$;

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

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

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

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

$$dx/dt = s(x, t), \qquad x(0) = x^0 . \tag{40}$$

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

$$x^{k+1} = T(x^k), \qquad T: \mathbf{R}^n \to \mathbf{R}^n , \tag{41}$$

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

$$V(T(x)) < V(x), \qquad x \neq T(x), \tag{42}$$

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

$$\|g(x) - g(y)\| \leq q\|x - y\|, \qquad q < 1, \tag{1}$$

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

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

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:

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

PROOF.

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

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

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

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

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 $g(x) = 0$, $g: \mathbf{R}^n \to \mathbf{R}^n$, 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

$$g(x) = y \tag{6}$$

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

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

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

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

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

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

and Newton's method

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

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 γ 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 γ
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

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

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

$$\gamma_k = \operatorname*{argmin}_{\gamma \geq 0} \phi_k(\gamma), \qquad \phi_k(\gamma) = f(x^k - \gamma \nabla f(x^k)). \tag{4}$$

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), $\nabla f(x^k) \to 0$ and the sequence x^k has limit points each of which is stationary, i.e., we can find the subsequence $x^{k_i} \to x^*$, 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. □

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

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

In (4), the γ_k can be written explicitly:

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

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

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

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

where ℓ 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 γ_k, we have

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

$$= f(x^k) - \frac{1}{2}\frac{\|\nabla f(x^k)\|^4}{(A\nabla f(x^k), \nabla f(x^k))}.$$

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*

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

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). □

Since

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

it follows from (7) that

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

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 γ. 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 $\gamma_k = 1/\lambda_{k+1}$, $k = 0, \ldots, n-1$, where λ_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 γ_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 γ can be advantageous. Let $0 < \varepsilon < 1$, $0 < \alpha < 1$ and let some γ be given. We compute $f(x^k - \gamma\nabla f(x))$ and verify the inequality

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

in each iteration. If it is satisfied, then $x^{k+1} = x^k - \gamma\nabla f(x^k)$; if not, then γ is replaced by γ_α 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 γ 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:

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

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

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

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

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

or the condition

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

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

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

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

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

For $\alpha_k = 0$ the method becomes Newton's; as $\alpha_k \to \infty$ the direction tends to the antigradient. Thus (16) is a compromise between these two methods. By appropriately choosing α_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

1. 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 γ in each iteration is finite, and method (11), with any rule (12)-(14) for choosing γ_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 $V(x) = \|\nabla f(x)\|^2$.

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^*.

3. 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 $x^k \to x^*$. Use the theorems of Section 2.2 and take $V(x) = f(x) - f(x^*)$.

4. 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 γ_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:

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

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

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

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

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

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 $f(x)$, $x \in \mathbf{R}^n$. Then for

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

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:

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

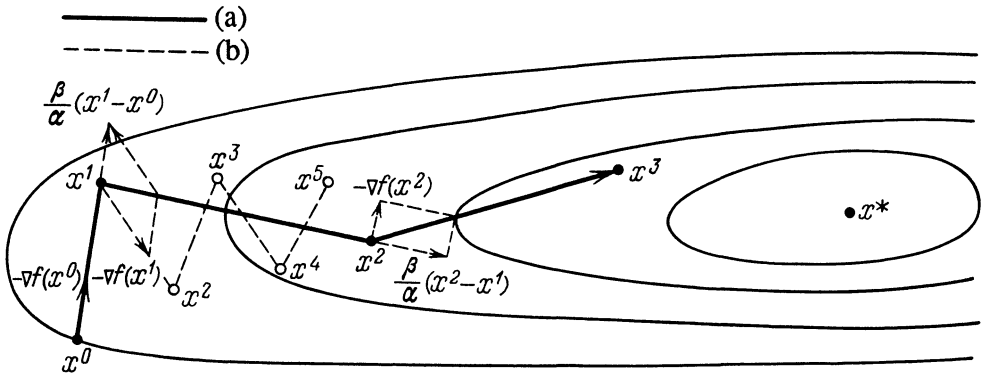


Fig. 6 (a) The heavy-ball method; (b) the gradient method.

The quantity q is minimal and equal to

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

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

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

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

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

Let $\ell = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n = L$ be the eigenvalues of the matrix B. Then

the eigenvalues ρ_j, $j = 1, ..., 2n$, of the matrix A coincide with the eigenvalues of 2×2-matrix of the form

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

Therefore, they are roots of the equations

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

One can show that if

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

then $|\rho| < 1$, where ρ 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 $\min_{\alpha,\beta} \max_{1 \le j \le 2n} |\rho_j|$, yields the optimal values α^*, β^* 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

$$q_1 = (L - \ell)/(L + \ell), \tag{10}$$

whereas for the two-step method it is equal to

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

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

$$q_1 \approx 1 - 2/\mu, \quad q_2 \approx 1 - 2/\sqrt{\mu}. \tag{12}$$

Hence, to be $e = 2.7,...$ 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 μ 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 α and β in (2) is not simple: we cannot directly use formulas (6), since the bounds of the spectrum of $\nabla^2 f(x^*)$ (the numbers ℓ and L) are usually unknown. □

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:

$$x^{k+1} = x^k - \alpha_k \nabla f(x^k) + \beta_k (x^k - x^{k-1}), \tag{13}$$

$$\{\alpha_k, \beta_k\} = \underset{\{\alpha,\beta\}}{\operatorname{argmin}} f(x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1})). \tag{14}$$

For a quadratic function

$$f(x) = (Ax, x)/2 - (b, x), \quad A > 0, \tag{15}$$

this problem can be solved explicitly:

$$\begin{aligned} \alpha_k &= \frac{\|r^k\|^2 (Ap^k, p^k) - (r^k, p^k)(Ar^k, p^k)}{(Ar^k, r^k)(Ap^k, p^k) - (Ar^k, p^k)^2}, \quad r^k = \nabla f(x^k) = Ax^k - b, \\ \beta_k &= \frac{\|r^k\|^2 (Ar^k, p^k) - (r^k, p^k)(Ar^k, r^k)}{(Ar^k, r^k)(Ap^k, p^k) - (Ar^k, p^k)^2}, \quad p^k = x^k - x^{k-1}. \end{aligned} \tag{16}$$

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 γ; 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^1) 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:

$$x^1 = x^0 - \frac{\|r^0\|^2}{(Ar^0, r^0)} r^0, \quad r^0 = \nabla f(x^0) = Ax^0 - b. \tag{17}$$

LEMMA 1. The gradients r^0, r^1, ... in each method, (13), (16), (17), are pairwise orthogonal:

$$(r^i, r^k) = 0, \quad i < k. \tag{18}$$

PROOF. We use induction on k. Let $(r^i, r^k) = 0$ for $0 \le i < k$, $k \ge 2$, and $r^i \ne 0$, $i = 0, ..., k$. 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 $r^i \ne 0$ for $i \le k$ that $\alpha_k \ne 0$. Hence Ar^k is a linear combination of r^{k+1}, r^k, and r^{k-1}, and similarly Ar^i, $i < k$, is a linear combination of r^{i+1}, r^i, r^{i-1}, and by induction, $(Ar^i, r^j) = 0$, $|i-j| > 1$, $i < k$, $j \le k$. Therefore

$$(r^{k+1}, r^i) = (r^k - \alpha_k A r^k + \beta_k (r^k - r^{k-1}), r^i) = 0 \quad \text{for } i = 0, ..., k-2.$$

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, ..., r^{k-1}, and r^{k-1} has the coefficient $-\alpha_{k-1} \ne 0$. Hence it follows from $(r^{k+1}, p^k) = 0$, $(r^{k+1}, r^i) = 0$, $i \le k-2$, that $(r^{k+1}, r^{k-1}) = 0$. Thus for all $i \le k$ one will have $(r^{k+1}, r^i) = 0$. □

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 $k \le n$ 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. □

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, ..., 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

$$(Ap^i, p^j) = 0, \quad i \neq j. \tag{19}$$

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 $r^0, \ldots, r^{k-1}$, $p^k = \sum_{j=0}^{k-1} \mu_j r^j$. 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, $i = 1, \ldots, n$, $(As^i, s^j) = 0$, $i \neq j$, allows us to solve easily the system

$$Ax = b, \quad A > 0. \tag{20}$$

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

$$\alpha_i = (b, s^i)/(As^i, s^i). \tag{21}$$

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 α_k are given by (21). Then $x^n = x^*$ is the solution of (20). Since the α_k in (21) can be determined differently: $\alpha_k = \arg\min_{\alpha} f(x^{k-1} + \alpha s^{k-1})$, 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

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k p^k , & \alpha_k &= \operatorname*{argmin}_{\alpha \geq 0} f(x^k + \alpha p^k) , \\ p^k &= -r^k + \beta_k p^{k-1} , & \beta_k &= \|r^k\|^2 / \|r^{k-1}\|^2 , \\ r^k &= \nabla f(x^k) , & \beta_0 &= 0 , \end{aligned} \tag{22}$$

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. □

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:

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k p^k , & \alpha_k &= \operatorname*{argmin}_{\alpha \geq 0} f(x^k + \alpha p^k) , \\ s^k &= -r^k + \beta_k p^{k-1} , & r^k &= \nabla f(x^k) , \end{aligned}$$

$$\beta_k = \begin{cases} 0 , & k = 0,n,2n,... \\ \|r^k\|^2 / \|r^{k-1}\|^2 , & k \neq 0,n,2n,... \end{cases} \tag{23}$$

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). □

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 β_k. The scheme

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k p^k , & \alpha_k &= \operatorname*{argmin}_{\alpha \geq 0} f(x^k + \alpha p^k) , \\ p^k &= -r^k + \beta_k p^{k-1} , & \beta_k &= \frac{(r^k,\ r^k - r^{k-1})}{\|r^{k-1}\|^2} , \\ r^k &= \nabla f(x^k) , & \beta_0 &= 0 , \end{aligned} \tag{24}$$

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,

$$\ell I \leq A \leq LI , \quad \ell > 0 , \tag{25}$$

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 $2(f(x^k) - f(x^*)) = (AP_k^2(A)(x^0 - x^*), x^0 - x^*) \le (AR^2(A)(x^0 - x^*), x^0 - x^*)$ 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 $r^0,\ldots,r^{k-1}$). Hence

$$\begin{aligned} \|x^k - x^*\|^2 &\le 2(f(x^k) - f(x^*))/\ell \le \|A\| \, \|R^2(A)\| \, \|x^0 - x^*\|^2/\ell \\ &\le (L/\ell)\|x^0 - x^*\| \max_{\ell \le \lambda \le L} R^2(\lambda) . \end{aligned} \tag{27}$$

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

$$R(\lambda) = T_k\left(\frac{L+\ell-2\lambda}{L-\ell}\right) \Big/ T_k\left(\frac{L+\ell}{L-\ell}\right) , \tag{28}$$

where $T_k(z)$ is the Chebyshev polynomial

$$T_k(z) = \begin{cases} [(z + \sqrt{z^2-1})^k + (z - \sqrt{z^2-1})^k]/2 , & |z| > 1 \\ \cos(k \arccos z), & |z| \le 1 . \end{cases} \tag{29}$$

Then

$$\begin{aligned} \max_{\ell \le \lambda \le L} R^2(\lambda) &= T_k^{-2}\left(\frac{L+\ell}{L-\ell}\right) \max_{-1 \le z \le 1} T_k^2(z) = T_k^{-2}\left(\frac{L+\ell}{L-\ell}\right) \\ &= 4(q^k + q^{-k})^{-2} \le 4q^{2k} , \qquad q = \frac{\sqrt{L} - \sqrt{\ell}}{\sqrt{L} + \sqrt{\ell}} . \end{aligned}$$

Hence

$$\|x^k - x^*\| \le 2\left(\frac{L}{\ell}\right)^{1/2} q^k \|x^0 - x^*\| , \qquad q = (\sqrt{L} - \sqrt{\ell})/(\sqrt{L} + \sqrt{\ell}) . \tag{30}$$

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 ℓ, 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 $\nabla f(x^k) \to 0$ (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 $\Sigma_{k=0}^{\infty} \varepsilon_k$, $\Sigma_{k=0}^{\infty} \varepsilon_k / (\varepsilon_0 + \cdots + \varepsilon_k)$ converge or diverge simultaneously), applying it to $\varepsilon_k = \|r^k\|^2/(\beta_1^2 \cdots \beta_k^2)$. 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 $\nabla f(x^k) \to 0$. 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:

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

where the matrix H_k is updated recursively on the basis of information obtained in the kth iteration, so that $H_k - [\nabla^2 f(x^k)]^{-1} \to 0$. 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

$$mI \leq H_k \leq MI, \quad m > 0. \tag{2}$$

Then in method (1) with $\gamma_k \equiv \gamma$, where $\gamma > 0$ is sufficiently small, one has $\nabla f(x^k) \to 0$. □

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

$$\|H_k - [\nabla^2 f(x^*)]^{-1}\| \to 0 . \tag{3}$$

Then method (1) with $\gamma_k = 1$ converges locally to x^* faster than any geometric progression. □

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 α_k along the coordinates and compute the gradients at these points. The corresponding difference approximation is the one sought if $\alpha_k \to 0$ (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

$$p^k = -H_k \nabla f(x^k) , \quad y^k = \nabla f(x^{k+1}) - \nabla f(x^k) . \tag{4}$$

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.,

$$\gamma_k p^k = A^{-1} y^k . \tag{5}$$

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*

$$H_{k+1} y^k = \gamma_k p^k \tag{6}$$

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 $a, b \in \mathbf{R}^n$. Also, let $(B^{-1}a, b) \neq -1$, and let $A = B + ab^T$. Then

$$A^{-1} = B^{-1} - (1 + (B^{-1}a, b))^{-1} B^{-1} a (B^{-1} b)^T. \quad (7)$$

The lemma is proved by straightforward verification. □

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):

$$H_{k+1} = H_k + \frac{\rho_k p^k (p^k)^T - p^k (y^k)^T H_k - H_k y^k (p^k)^T}{(y^k, p^k)},$$

$$\rho_k = \gamma_k + \frac{(H_k y^k, y^k)}{(y^k, p^k)}, \quad H_0 > 0. \quad (10)$$

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 $\gamma_k = \operatorname{argmin}_{\gamma} f(x^k + \gamma p^k)$ for $f(x) = (Ax,x)/2 - (b,x)$, $A > 0$, is finite: $x^n = x^* = A^{-1}b$. □

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

1. Let $e_1, \ldots, e_n$ 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 ith row. Show that $H(\alpha) \to \nabla^2 f(x)$ as $\alpha \to 0$.

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:

$$(x, y)_1 = (Ax, y) . \tag{11}$$

In this case the A defines a new metric in $\mathbf{R}^n$:

$$\|x - y\|_1^2 = (A(x-y), x - y) . \tag{12}$$

Let us write the gradient of a differentiable function $f(x)$ in the new metric:

$$f(x+y) = f(x) + (\nabla f(x), y) + o(\|y\|) = f(x) + (AA^{-1}\nabla f(x), y) + o(\|y\|)$$

$$= f(x) + (a, y)_1 + o(\|y\|_1) ,$$

$$a = A^{-1}\nabla f(x) .$$

By definition, the vector a is the gradient of $f(x)$ in space with scalar product (11). Thus,

$$\nabla_1 f(x) = A^{-1}\nabla f(x) . \tag{13}$$

In the new metric the gradient method assumes the form

$$x^{k+1} = x^k - \gamma_k \nabla_1 f(x^k) = x^k - \gamma_k A^{-1}\nabla f(x^k) \tag{14}$$

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

$$f(x) = (Bx, x)/2 - (b, x) = (\tfrac{1}{2})(A^{-1}Bx, x)_1 - (A^{-1}b, x)_1 \tag{15}$$

the convergence rate of (14) is determined by the progression ratio $q = (L - \ell)(L + \ell)$, where L and ℓ 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 μ is equal to one.

For a nonquadratic function the method

$$x^{k+1} = x^k - \gamma_k H_k \nabla f(x^k) , \qquad H_k > 0 , \tag{16}$$

can be viewed as the gradient method in the metric

$$(x, y)_1 = (H_k^{-1}x, y) , \tag{17}$$

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 approch 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 $p^0,...,p^{n-1}$:

$$(Ap^i, p^j) = 0, \quad i \neq j, \tag{18}$$

makes it possible to find the minimum of a quadratic function $f(x) = (Ax, x)/2 - (b, x)$ in n one-dimensional minimizations:

$$x^{k+1} = x^k + \alpha_k p^k, \quad \alpha_k = \operatorname*{argmin}_{\alpha} f(x^k + \alpha p^k). \tag{19}$$

Then $x^n = x^* = A^{-1}b$ 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 $p^1, ..., p^k$, $k < n-1$ be conjugate vectors that have been constructed,

$$(Ap^i, p^j) = 0, \quad 0 \leq i, j \leq k, \ i \neq j, \tag{20}$$

and let x^k be the corresponding points in method (19). The next vector p^{k+1} must satisfy the relation

$$(p^{k+1}, Ap^i) = 0, \quad i = 0,...,k.$$

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

$$(p^{k+1}, y^i) = 0, \quad i = 1,...,k. \tag{21}$$

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:

$$x^{k+1} = x^k + \alpha_k p^k , \quad \alpha_k = \operatorname*{argmin}_{\alpha \geq 0} f(x^k + \alpha p^k)$$
$$(p^{k+1}, y^i) = 0 , \quad i = 1,...,k , \quad y^i = \nabla f(x^{i+1}) - \nabla f(x^i) . \tag{22}$$

Usually, p^{k+1} is sought here in the form

$$p^{k+1} = -H_{k+1} \nabla f(x^{k+1}) , \quad H_{k+1} = H_k + \Delta H_k \tag{23}$$

and the matrix H_k is stored instead of the vectors y^i, $i = 1,...,k$. The methods thus assume the same form (1) as the quasi-Newton methods. The only difference is that it is not necessarily $H_k \to [\nabla^2 f(x^k)]^{-1}$; 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:

$$x^{k+1} = x^k + \alpha_k p^k , \qquad \alpha_k = \operatorname*{argmin}_{\alpha \geq 0} f(x^k + \alpha p^k) ,$$
$$p^k = -H_k \nabla f(x^k) , \qquad y^k = \nabla f(x^{k+1}) - \nabla f(x^k) ,$$
$$H_{k+1} = H_k - \frac{H_k y^k (y^k)^T H_k}{(H_k y^k, y^k)} , \quad k+1 \neq n, 2n, \ldots , \tag{24}$$
$$H_0 = H_n = H_{2n} = \cdots = I .$$

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 minimim point has been proved.

3.3.3 The Secant Method

One the simplest and most commonly used methods for solving the one-dimensional equation

$$g(x) = 0 \tag{25}$$

is the secant method illustrated in Figure 7. This method can be extended to the multidimensional case: if g: $\mathbf{R}^n \to \mathbf{R}^n$, 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).

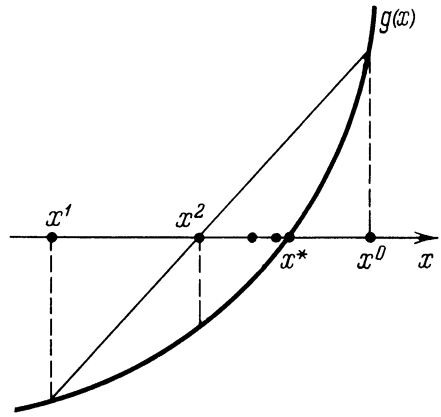


Fig. 7 The secant method.

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}, ..., x^{k-n} be $n+1$ points in $\mathbf{R}^n$, and let $\nabla f(x^k)$, ..., $\nabla f(x^{k-n})$ be the gradients computed at these points. We solve the system of $n+1$ linear equations with $n+1$ variables $\lambda_0, \lambda_1, ..., \lambda_n$:

$$\sum_{i=0}^{n} \lambda_i \nabla f(x^{k-i}) = 0, \quad \sum_{i=0}^{n} \lambda_i = 1. \tag{26}$$

Also, let us construct the point

$$x^{k+1} = \sum_{i=0}^{n} \lambda_i x^{k-i}. \tag{27}$$

Then the process is repeated for the last $n+1$ points x^{k+1}, x^k, ..., 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, ..., 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 $f(x)$. □

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 $b^1, \ldots, b^n$. Also, let $\tilde{B}$ differ from it only in the first column (say, b^1 is replaced by $\tilde{b}^1$). Then

$$\tilde{c}^i = c^i - \frac{(\tilde{b}^1 - b^1, c^i)}{1 + (\tilde{b}^1 - b^1, c^1)} c^1, \tag{28}$$

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 $e = (1, 0, \ldots, 0)$, and use Lemma 3. □

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 $\Sigma_i \lambda_i x^{k-i}$). 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

$$f(x^* + \lambda(x-x^*)) - f(x^*) = \lambda^\gamma (f(x) - f(x^*)) \tag{29}$$

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

$$f(x) - f(x^*) = \gamma^{-1} (\nabla f(x), x-x^*) . \tag{30}$$

To prove (30) we take $\lambda = 1 + \varepsilon$ in (29):

$$f(x + \varepsilon(x-x^*)) - f(x^*) = (1+\varepsilon)^\gamma (f(x) - f(x^*)) ,$$

$$\varepsilon\gamma(f(x) - f(x^*)) = \varepsilon(\nabla f(x), x-x^*) + o(\varepsilon) .$$

Letting ε 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 γ is known, then taking $n+1$ points $x^0, \ldots, x^n$, yields the system

$$\gamma f(x^i) - \alpha + (\nabla f(x^i), x^*) = (\nabla f(x^i), x^i) , \quad i = 0,\ldots,n , \tag{31}$$

which is linear in the $n+1$ variables x^*, $\alpha(\alpha = \gamma f(x^*))$. Eliminating α, we obtain n linear equations to determine $x^* \in \mathbf{R}^n$:

$$(\nabla f(x^i) - \nabla f(x^0), x^*) = (\nabla f(x^i), x^i) - (\nabla f(x^0), x^0) - \gamma(f(x^i) - f(x^0)) ,$$

$$i = 1,\ldots,n . \tag{32}$$

But if γ is unknown, then one can take $n+2$ points $x^0, \ldots, x^{n+1}$, and determine the $n+1$ variables γ, 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 $x^0, \ldots, x^k$, $k > n$ have been constructed. Taking the last $n+1$ among them (or $n+2$) if γ is unknown), we solve the system (with respect to x, α, γ, or x, α)

$$(\nabla f(x^i) - \alpha + \gamma f(x^i) = (\nabla f(x^i), x^i), \quad i = k, k-1, \ldots, \tag{33}$$

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 α/γ) 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 γ being known. We consider the gradient method

$$x^{k+1} = x^k - \frac{\gamma(f(x^k) - f^*)}{\|\nabla f(x^k)\|^2} \nabla f(x^k). \tag{34}$$

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

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - \frac{2\gamma(f(x^k) - f^*)}{\|\nabla f(x^k)\|^2}(\nabla f(x^k), x^k - x^*) + \frac{\gamma^2 (f(x^k) - f^*)^2}{\|\nabla f(x^k)\|^2}$$

$$= \|x^k - x^*\|^2 - \frac{\gamma^2 (f(x^k) - f^*)^2}{\|\nabla f(x^k)\|^2}$$

implying that if $\|\nabla f(x)\|$ is bounded on the set $\{x: \|x - x^*\| \leq \|x^0 - x^*\|\}$, then $f(x^k) \to f^*$. It is not hard to see that this result still holds if in (30) equality is replaced by inequality

$$f(x) - f^* \leq \gamma^{-1}(\nabla f(x), x - x^*). \tag{35}$$

A somewhat different class (versus the homogeneous one) is given by the formula

$$f(x) = F(\phi(x)), \quad \phi(x) = (Ax, x)/2 - (b, x), \quad A > 0, \tag{36}$$

where $F: \mathbf{R}^1 \to \mathbf{R}^1$ is a monotone function on $[\phi^*, \infty]$, $\phi^* = \phi(x^*)$. Obviously, x^* is a minimum point of $f(x)$.

If F and ϕ 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:

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k p^k, \qquad \alpha_k = \operatorname*{argmin}_{\alpha \geq 0} f(x^k + \alpha p^k), \\ p^k &= -\nabla f(x^k) + \beta_k p^{k-1}, \\ \beta_k &= \frac{F'(\phi(x^{k-1}))\ \|\nabla f(x^k)\|^2}{F'(\phi(x^k))\ \|\nabla f(x^{k-1})\|^2}, \quad \beta_0 = 0. \end{aligned} \tag{37}$$

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 β_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 $(a^i, x) = \beta_i$, $i = 1,...,m$, $x \in \mathbf{R}^n$. 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 γ.

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 2α.

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: $\mathbf{R}^n \to \mathbf{R}^1$ 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

$$|\Delta_1 - (\nabla f(x), y)| \to 0 \quad \text{as} \quad \alpha \to 0. \quad (2)$$

(b) If ∇f satisfies a Lipschitz condition with constant L in a neighborhood of x, then for a sufficiently small α

$$|\Delta_1 - (\nabla f(x), y)| \leq L\alpha\|y\|^2/2. \quad (3)$$

(c) If f is twice differentiable and $\nabla^2 f$ satisfies a Lipschitz condition in a neighborhood of x, then for a sufficiently small α

$$|\Delta_2 - (\nabla f(x), y)| \leq L\alpha^2 \|y\|^3/6 . \tag{4}$$

(d) If $f(x)$ is quadratic, then for any α

$$\Delta_2 = (\nabla f(x), y) . \tag{5}$$

Lemma 1 is easily proved if one uses (2), (15), (20) of Section 1.1. □

Thus, the difference relations Δ_1 and Δ_2 may serve as an approximation for linear approximation of $f(x)$. Let us consider methods of the form

$$x^{k+1} = x^k - \gamma_k s^k , \tag{6}$$

where $\gamma_k \geq 0$ is the step size and s^k is computed by one of these two formulas:

$$s^k = \sum_{i=1}^{m} \alpha_k^{-1} [f(x^k + \alpha_k h^i) - f(x^k)] h^i , \tag{7}$$

$$s^k = \sum_{i=1}^{m} (2\alpha_k)^{-1} [f(x^k + \alpha_k h^i) - f(x^k - \alpha_k h^i)] h^i , \tag{8}$$

where h^i, $i = 1,...,m$, are vectors giving the directions of the trial steps, α_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: $m = n$, $h^i = e_i$, $i = 1,...,n$, 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:

$$x_i^{k+1} = x_i^k - (\gamma_k/\alpha_k)[f(x^k + \alpha_k e_i) - f(x^k)] . \tag{9}$$

By Lemma 1

$$s^k = \sum_{i=1}^{n} (\nabla f(x^k), e_i) e_i + \varepsilon^k = \nabla f(x^k) + \varepsilon^k , \tag{10}$$

where the remainder ε^k can be estimated either for (7) or for (8), depending on the smoothness of $f(x)$.

(b) Method of coordinatewise descent: $m = 1$, $h = e_j$, $j = k(\mathrm{mod}\ n)$.

The steps are made along the coordinates chosen in cyclic order:

$$x_i^{k+1} = \begin{cases} x_i^k - (\gamma_k/\alpha_k)\,[\,f(x^k + \alpha_k e_i) - f(x^k)\,], & i = k \pmod n, \\ x_i^k \quad \text{otherwise}. \end{cases} \tag{11}$$

Here $s^k = \nabla f(x^k)_j e_j + \varepsilon^k$.

(c) Method of random coordinatewise descent: $m = 1$, $h = e_j$, where j takes on the values $1,\ldots,n$ 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

$$x^{k+1} = x^k - (\gamma_k/\alpha_k)[f(x^k + \alpha_k h) - f(x^k)]h\,. \tag{12}$$

The convergence of all the methods is guaranteed by the condition $\alpha_k \to 0$ (see Exercise 1).

The rate of convergence depends on the smoothness of $f(x)$ and the choice of α_k. To minimize errors, it is more convenient to take large α_k since the smaller the α_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 α_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 $\alpha_k \le cq^k$, where q ~~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:

$$f(x) = (Ax, x)/2 - (b, x), \qquad A > 0\,, \tag{13}$$

and let γ_k be chosen from the steepest descent condition:

$$x^{k+1} = x^k - \gamma_k s^k\,, \qquad \gamma_k = \operatorname*{argmin}_{\gamma \ge 0} f(x^k - \gamma s^k)\,. \tag{14}$$

We shall compare three methods of choosing the s^k: (1) the symmetric difference approximation of the gradient

$$s^k = \sum_{i=1}^{n} (2\alpha)^{-1}[f(x^k + \alpha e_i) - f(x^k - \alpha e_i)]e_i = \nabla f(x^k) \tag{15}$$

(the last equality is due to (5)); (2) coordinatewise descent

$$s^k = (2\alpha)^{-1}[f(x^k+\alpha e_i) - f(x^k-\alpha e_i)]e_i = \nabla f(x^k)_i e_i\,, \quad i = k(\mathrm{mod}\ n) \quad (16)$$

and (3) random search

$$s^k = (2\alpha)^{-1}[f(x^k+\alpha h^k) - f(x^k-\alpha h^k)]h^k = (\nabla f(x^k), h^k)h^k\,, \quad (17)$$

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 γ_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 $\alpha_k \to 0$, $\gamma_k = \gamma$, where γ is sufficiently small, one can assert for all the methods a-d that $\nabla f(x^k) \to 0$ a.s. Use the technique employed in proving Theorem 1 of Section 2.2.

2. Suggest a constructive method for regulating α_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 $x^0, x^1, \ldots, x^n$, 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: $j = \underset{0 \le i \le n}{\operatorname{argmax}} f(x^i)$. Next, we construct a new simplex different from the old one only in one vertex: x^j is replaced by x^{n+1}:

$$x^{n+1} = 2n^{-1}(x^0 + \cdots + x^{j-1} + x^{j+1} + \cdots + x^n) - x^j \qquad (18)$$

(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).

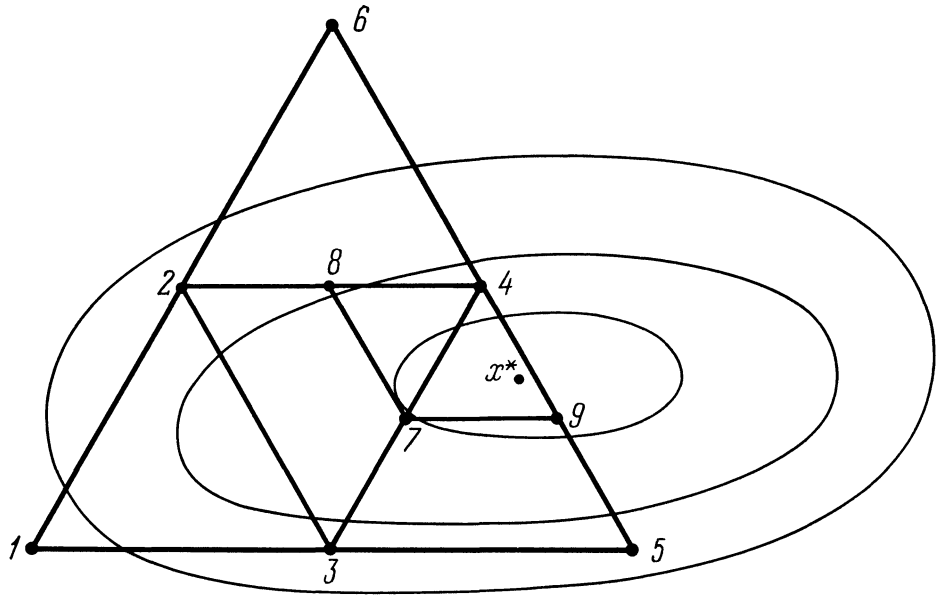


Fig. 8 The simplicial method.

In addition to this simplest variant of the simplicial method, there exist many modications 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 $x^0, \ldots, x^n$. Then one computes the values of the $f(x)$ at all of these points and at the midpoints on the segments joining them (let $f((x^i+x^j)/2) = f_{ij}$, $f(x^i) = f_{ii}$, $i,j = 0,\ldots,n$). Lastly, one solves the system of linear (with respect to $\lambda, \lambda_0, \ldots, \lambda_n$) equations

$$\begin{aligned} 4\sum_{j=0}^{n} f_{ij}\lambda_j + \lambda &= f_{ii}, \quad i = 0,\ldots,n, \\ \sum_{j=0}^{n} \lambda_i &= 1 \end{aligned} \tag{19}$$

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* $x^0, \ldots, x^n$ such that $x^n - x^0, \ldots, x^1 - x^0$ 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 λ_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 aproximation. 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 $f_{ij} \le (f_{ii} + f_{jj})/2$, 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.

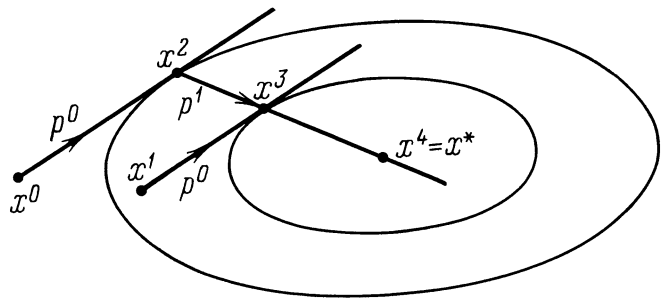


Fig. 9 The method of conjugate directions.

LEMMA 2. Let $f(x) = (Ax, x)/2 - (b, x)$, $A > 0$, $x \in \mathbf{R}^n$, $p^1, \ldots, p^k$ be conjugate vectors:

$$(Ap^i, p^j) = 0, \qquad i \ne j, \quad k < n,$$

$$L^0 = \left\{x: x = x^0 + \sum_{i=1}^{k} \lambda_i p^i\right\}, \qquad x^1 \in L^0,$$

$$L^1 = \left\{x: x = x^1 + \sum_{i=1}^{k} \lambda_i p^i\right\},$$

$$y^0 = \operatorname*{argmin}_{x \in L^0} f(x), \qquad y^1 = \operatorname*{argmin}_{x \in L^1} f(x).$$

Then the vector $p^{k+1} = y^1 - y^0$ is conjugate to $p^1, \ldots, p^k$.

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). □

One can thus construct a minimization method in the following way. Let x^k be the approximation to the solution obtained in the kth iteration, and let $p^0, \ldots, p^k$ 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 $p^0, \ldots, p^k$. Now let us make sequential one-dimensional minimizations in the directions $p^0, \ldots, p^k$, starting at the point $\bar{x}^k$; we obtain the point $\tilde{x}^{k+1}$. For the point x^{k+1} we take the minimum point of $f(x)$ on the line joining $\tilde{x}^{k+1}$ with x^k, and for p^{k+1} we take the vector $\tilde{x}^{k+1} - x^k$. 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 ε can be usually assumed to be constant (not depending on x^k) and

generally not too large. If necessary, ε 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 statitistical 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:

$$f(x) = \mathbf{E}Q(x, \omega) = \int Q(x, \omega)\, d\mathbf{P}(\omega) , \qquad (1)$$

where the function $Q(x, \omega)$ is known but the distribution $\mathbf{P}(\omega)$ is not specified. Only a sample $\omega_1, \ldots, \omega_k$ 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

$$\frac{1}{k}\sum_{i=1}^{k} Q(x, \omega_i) \quad \text{and} \quad \frac{1}{k}\sum_{i=1}^{k} \nabla_x Q(x, \omega_i) , \qquad (2)$$

or more simply

$$Q(x, \omega_k) \quad \text{and} \quad \nabla_x Q(x, \omega_k) . \qquad (3)$$

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

$$s^k = \nabla f(x^k) + r^k , \tag{4}$$

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

$$\|r^k\| \leq \varepsilon , \tag{5}$$

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

$$\|r^k\| \leq \varepsilon \|\nabla f(x^k)\| . \tag{6}$$

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:

$$\mathbf{E} r^k = 0 , \quad \mathbf{E}\|r^k\| \leq \sigma^2 . \tag{7}$$

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:

$$\mathbf{E} r^k = 0 , \quad \mathbf{E}\|r^k\|^2 \leq \tau \|\nabla f(x^k)\|^2 . \tag{8}$$

Of course, other types of noise, too, are observed in practice; for example, random noise with systematic error ($\|\mathbf{E}r^k\| \le \varepsilon$) or random bounded noise ($\mathbf{E}r^k = 0$, $\|r^k\| \le \varepsilon$). 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 ε_k depends on the number of the iteration and that $\varepsilon_k \to 0$ as $k \to \infty$. 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:

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

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 ℓ) 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 $\|r^k\| \le \varepsilon$, $\gamma_k \equiv \gamma$. Then we can find a $\bar{\gamma} > 0$ such that for $0 < \gamma < \bar{\gamma}$ in method (1) we have

$$\|x^k - x^*\| \le \rho + q^k \|x^0 - x^*\|, \tag{2}$$

where $0 \le q < 1$, $\rho = \rho(\varepsilon) \to 0$ as $\varepsilon \to 0$, x^* is the minimum point of $f(x^k)$.

PROOF. Introduce the Lyapunov function

$$V(x) = \frac{1}{2}\left(\|x - x^*\| - \frac{1}{\ell}\varepsilon\right)_+^2. \tag{3}$$

Using the result of Exercise 1 below, we obtain

$$(\nabla V(x^k), s^k) = \left(\|x^k - x^*\| - \frac{1}{\ell}\varepsilon\right)_+ \frac{(\nabla f(x^k)+r^k,\ x^k-x^*)}{\|x^k - x^*\|}$$

$$\geq \left(\|x^k - x^*\| - \frac{1}{\ell}\varepsilon\right)_+ (\ell\|x^k - x^*\| - \varepsilon) = 2\ell V(x^k),$$

$$\|s^k\|^2 = \|\nabla f(x^k) + r^k\|^2 \leq (L\|x^k - x^*\| + \varepsilon)^2 \leq a + bV(x^k)$$

$$\leq a + (b/(2\ell))(\nabla V(x^k), s^k),$$

where a and b are constants, and $a \to 0$ as $\varepsilon \to 0$. Applying Theorem 4 of Section 2.2 proves the theorem. □

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 γ 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 ρ, γ, 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

$$f(x) = (Ax,x)/2 - (b,x), \quad \ell I \leq A \leq LI, \ \ell > 0, \ \|r^k\| \leq \varepsilon, \ 0 < \gamma < 2/L.$$

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

$$\|x^k - x^*\| \leq \gamma\varepsilon/(1-q) + q^k(\|x^0 - x^*\| - \gamma\varepsilon/(1-q)).$$

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

$$\|r^k\| \leq \alpha\|\nabla f(x^k)\| , \quad \alpha < 1, \ \gamma_k \equiv \gamma .$$

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

$$(\nabla V(x), s^k) = (\nabla f(x^k), \nabla f(x^k) + r^k) \geq (1 - \alpha)\|\nabla f(x^k)\|^2$$

$$\geq (1 - \alpha)2\ell V(x^k) ,$$

$$\|s^k\|^2 \leq \|\nabla f(x^k)\|^2(1 + \alpha)^2 \leq 2(1 + \alpha)^2 L V(x^k) .$$

It remains only to apply Theorem 4 of Section 2.2. □

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 $\mathbf{E}r^k = 0$ and $\mathbf{E}\|r^k\| \leq \sigma^2$.

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

$$\mathbf{E}(f(x^k) - f^*) \leq \rho(\gamma) + \mathbf{E}(f(x^0) - f^*)q^k , \qquad (4)$$

where $q < 1$, $\rho(\gamma) \to 0$ as $\gamma \to 0$. If

$$\gamma_k \to 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty, \tag{5}$$

then $\mathbf{E}\|x^k - x^*\|^2 \to 0$. If though

$$\sum_{k=0}^{\infty} \gamma_k^2 < \infty, \quad \sum_{k=0}^{\infty} \gamma_k = \infty, \tag{6}$$

then $x^k \to x^*$ a.s. Finally, if $\gamma_k = \gamma/k$, $\gamma > (2\ell)^{-1}$, then

$$\mathbf{E}(f(x^k) - f^*) \leq \frac{L\sigma^2\gamma^2}{2(2\ell\gamma - 1)k} + o\left(\frac{1}{k}\right). \tag{7}$$

PROOF. Take $V(x) = f(x) - f^*$. Then

$$(\nabla V(x^k), \mathbf{E}s^k) = (\nabla f(x^k), \nabla f(x^k)) \geq 2\ell V(x^k),$$

$$\mathbf{E}\|s^k\|^2 = \|\nabla f(x^k)\|^2 + \mathbf{E}\|r^k\|^2 \leq \sigma^2 + (\nabla V(x^k), \mathbf{E}s^k).$$

It remains only to use Theorems 2-5 of Section 2.2. □

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 γ, the smaller the size of this region. Secondly, choosing decreasing γ_k may make the method converge in some probabilistic sense (in the mean as $\gamma_k \to 0$ or almost surely for $\Sigma_{k=0}^{\infty} \gamma_k^2 < \infty$). Thirdly, the convergence rate is quite slow in this case (of order $O(1/k)$). As will be shown later, no choice of γ_k can yield a better convergence rate.

Let us refine Theorem 3 for a quadratic function and constant noise level. Thus, let

$$f(x) = (Ax, x)/2 - (f, x), \quad \ell I \leq A \leq LI, \quad \ell > 0, \tag{8}$$

$$\mathbf{E}r^k = 0, \quad \mathbf{E}r^k(r^k)^T = \sigma^2 I.$$

We assume that the initial approximation x^0 is random and symmetrically distributed around x^*: $\mathbf{E}(x^0 - x^*)(x^0 - x^*)^T \neq \alpha I$.

THEOREM 4. For any $0 < \gamma < 2/L$, $\gamma_k \equiv \gamma$, in method (1) under conditions (8) for the quantity

$$U_k = \mathbf{E}(x^k - x^*)(x^k - x^*)^T \tag{9}$$

we have the relations

$$U_k \to U_\infty = \gamma\sigma^2 A^{-1}(2I - \gamma A)^{-1}, \tag{10}$$

$$\|U_k - U_\infty\| \le \|U_0 - U_\infty\| q^k, \tag{11}$$

$$q = \max\{(1-\gamma\ell)^2, (1-\gamma L)^2\} < 1.$$

If $\gamma_k \equiv \gamma/k$, $\gamma > (2\ell)^{-1}$

$$U_k = \frac{1}{k}B(\gamma) + o\left(\frac{1}{k}\right), \quad B(\gamma) = \gamma\sigma^2\left(2A - \frac{1}{\gamma}I\right)^{-1}. \tag{12}$$

The quantity $\|B(\gamma)\|$ is minimal for $\gamma = 1/\ell$,

$$\|U_k\| = \frac{1}{k}\frac{\sigma^2}{\ell^2} + o\left(\frac{1}{k}\right). \quad \square \tag{13}$$

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

$$\mathbf{E}\|r^k\|^2 \le \alpha\|\nabla f(x)\|^2. \tag{14}$$

THEOREM 5. For any α we can find a $\bar{\gamma}$ such that for $0 < \gamma < \bar{\gamma}$, in method (1) we have

$$\mathbf{E}\|x^k - x^*\|^2 \le cq^k, \quad q < 1. \quad \square \tag{15}$$

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:

$$s^k = [\nabla^2 f(x^k)]^{-1} \nabla f(x^k) + r^k , \tag{1}$$

where r^k is the noise, and the step is

$$x^{k+1} = x^k - s^k . \tag{2}$$

Suppose that the noise contains a systematic error:

$$\|r^k\| \leq \varepsilon . \tag{3}$$

As is known, Newton's method converges locally in some region U. If ε 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 μ 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

$$f(x) = (Ax, x)/2 - (b, x), \quad A > 0, \tag{4}$$

where A and A^{-1} are computed exactly, and where the gradient contains a random error:

$$s^k = \nabla f(x^k) + r^k = Ax^k - b + r^k, \quad \mathbf{E}r^k = 0, \ \mathbf{E}\|r^k\|^2 = \sigma^2 . \tag{5}$$

Consider the method

$$x^{k+1} = x^k - \gamma_k A^{-1} s^k , \tag{6}$$

a generalization of Newton's method, by introduction of the parameter γ_k. As we will see later (Theorem 1, Section 4.5), this method cannot converge faster than $O(1/k)$ for any γ_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

$$\|r^k\| \leq c\|\nabla f(x^k)\|^2 , \tag{7}$$

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

$$\begin{aligned} &f(x) = (Ax, x)/2 - (b, x), && \ell I \leq A \leq LI, \ \ell > 0 , \\ &s^k = \nabla f(x^k) + r^k = Ax^k - b + r^k , && \mathbf{E}r^k = 0, \ \mathbf{E}r^k(r^k)^T = \sigma^2 I , \end{aligned} \tag{8}$$

where the r^k are mutually independent. As can be shown, the heavy-ball method with constant coefficients

$$x^{k+1} = x^k - \alpha s^k + \beta(x^k - x^{k-1}) \tag{9}$$

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

$$x^{k+1} = x^k - \alpha_k y^k, \quad y^{k+1} = y^k - \beta_k(y^k - s^k). \tag{10}$$

At the same time, let us consider the gradient method

$$x^{k+1} = x^k - \gamma_k s^k, \tag{11}$$

limiting coefficients to the form

$$\alpha_k = \frac{1}{k}\alpha, \quad \beta_k = \frac{1}{k}\beta, \quad \gamma_k = \frac{1}{k}\gamma. \tag{12}$$

THEOREM 1. For any α, β, 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 $\gamma_k = 1/k\ell\Gamma$ □

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:

$$Ap^i \approx y^i, \quad p^i = x^{i+1} - x^i, \quad y^i = \nabla f(x^{i+1}) - \nabla f(x^i), \quad i = 0,\ldots,k-1. \tag{13}$$

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 η_k. As above, we will speak of an absolute (relative) deterministic error if $|\eta_k| \le \varepsilon$ ($|\eta_k| \le \alpha(f(x^k) - f(x^*))$), and of an absolute (relative) random error if the η_k are random, independent,

$$\mathbf{E}\eta_k = 0 \quad \text{and} \quad \mathbf{E}\eta_k^2 \le \sigma^2 \quad (\mathbf{E}\eta_k^2 \le \tau(f(x^k) - f(x))).$$

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):

$$x^{k+1} = x^k - \gamma_k s^k, \quad s^k = \sum_{i=1}^{n} \frac{1}{2\alpha_k}(\tilde{f}(x^k + \alpha_k e_i) - \tilde{f}(x^k - \alpha_k e_i))e_i, \tag{1}$$

where e_i are the standard basis vectors. Here and later

$$\tilde{f}(x) = f(x) + \eta , \tag{2}$$

and the random errors η are independent at different points and

$$\mathbf{E}\eta = 0 , \quad \mathbf{E}\eta^2 \leq \sigma^2 . \tag{3}$$

Let us discuss the trial and working steps α_k, γ_k. Set

$$s^k - \nabla f(x^k) = g^k + \xi^k ,$$

where g^k is the systematic error, and ξ^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,

$$\|g^k\| \leq c\alpha_k^2 . \tag{4}$$

The random component of the error in estimating the gradient is:

$$\mathbf{E}\xi^k = 0 , \quad \mathbf{E}\|\xi^k\|^2 \leq \sigma^2/(2\alpha_k^2) . \tag{5}$$

Thus the systematic error decreases as α_k decreases, but the random error increases. First let us show that α_k, γ_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 γ_k, α_k let the following relations hold:

$$\sum_{k=0}^{\infty} \gamma_k = \infty , \quad \sum_{k=0}^{\infty} \gamma_k \alpha_k^4 < \infty , \quad \sum_{k=0}^{\infty} \gamma_k^2 \alpha_k^2 < \infty ,$$
$$\sum_{k=0}^{\infty} \gamma_k^2 \alpha_k^{-2} < \infty . \tag{6}$$

Then in method (1), $x^k \to x^*$ a.s. and $\mathbf{E}\|x^k - x^*\|^2 \to 0$. If $\gamma_k = \gamma/k$, $\alpha_k = \alpha k^{-1/6}$ and γ is sufficiently large, then

$$\mathbf{E}\|x^k - x^*\|^2 = O(k^{-2/3}) . \quad \square$$

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 α_k, γ_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:

$$\begin{aligned} f(x) &= (Ax, x)/2 - (b, x), \qquad A \geq \ell I > 0, \ x \in \mathbf{R}^n, \\ \tilde{f}(x) &= f(x) + \eta, \qquad\qquad \mathbf{E}\eta = 0, \ \mathbf{E}\eta^2 = \sigma^2, \end{aligned} \tag{7}$$

where the noise η 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 η). Since for a quadratic function the systematic error in the difference approximation of the gradient is equal to 0 for any α_k (Lemma 1 of Section 3.4), it is not necessary here to make α_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 γ_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 α_k.

Exercises

1. Show that among the α_k, γ_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 $\mathbf{E}\|x^k - x^*\|^2 = O(1/k^{1/2})$.

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: $y_i = (a,x^i) + \beta + \eta_i$, $i = 1,...,k$, where η_i is random independent noise, $\mathbf{E}\eta_i = 0$, $\mathbf{E}\eta_i^2 = \sigma^2$. 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.,

$$c^k = \operatorname*{argmin}_c \sum_{i=1}^{k} (y_i - (c,z^i))^2 = \left(\sum_{i=1}^{k} z^i (z^i)^T\right)^{-1} \left(\sum_{i=1}^{k} z^i y_i\right) = \Gamma_k \sum_{i=1}^{k} z^i y_i ,$$

$$\Gamma_k = \left(\sum_{i=1}^{k} z^i (z^i)^T\right)^{-1} . \tag{12}$$

This method can be given a recursive form—the new measurement at the point x^{k+1}:

$$y_{k+1} = (c^*, z^{k+1}) + \eta_{k+1}, \qquad z^{k+1} = \{x^{k+1}, 1\},$$

can be taken into account by means of the formula

$$c^{k+1} = c^k - \Gamma_{k+1} z^{k+1} ((c^k, z^{k+1}) - y_{k+1}) ,$$

$$\Gamma_{k+1} = \Gamma_k - \frac{\Gamma_k z^{k+1} (\Gamma_k z^{k+1})^T}{1 + (\Gamma_k z^{k+1}, z^{k+1})}, \qquad k \geq n+1 , \tag{13}$$

$$\Gamma_{k+1} = \left(\sum_{i=1}^{n+1} z^i (z^i)^T\right)^{-1} .$$

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 inlcude "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 $\mathbf{R}^n$:

$$x^{k+1} = x^k - \gamma_k s^k , \qquad s^k = R(x^k) + \xi^k , \tag{1}$$

where $\gamma_k \geq 0$ are deterministic scalar factors, $R(x)$ is some function, and ξ^k are random noises assumed to be independent and centered ($\mathbf{E}\xi^k = 0$). The initial approximation x^0 can either be deterministic or random, and in the latter case it is assumed that $\mathbf{E}\|x^0\|^2 < \infty$ and x^0, ξ^i are independent. Suppose that there exists a unique point x^* such that $R(x^*) = 0$ and $R(x)$ satisfies the linear growth condition:

$$\|R(x)\| \leq L\|x - x^*\| . \tag{2}$$

THEOREM 1. For all k let

$$\mathbf{E}\|\xi^k\|^2 \geq \sigma^2 . \tag{3}$$

Then under the assumptions made above, for any method (1)

$$\mathbf{E}\|x^k - x^*\|^2 \geq 1/(a + kb) , \qquad a = 1/\mathbf{E}\|x^0 - x^*\|^2, \ b = L^2/\sigma^2 . \tag{4}$$

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 γ_k, in particular to a choice where convergence fails to occur.

PROOF. Let us estimate the conditional mathematical expectation $\mathbf{E}(\|x^{k+1} - x^*\|^2 \mid x^k)$:

$$\mathbf{E}(\|x^{k+1} - x^*\|^2 \mid x^k) = \|x^k - x^* - \gamma_k R(x^k)\|^2 + \gamma_k^2 \mathbf{E}\|\xi^k\|^2 ,$$

$$\|x^k - x^* - \gamma_k R(x^k)\| \geq (\|x^k - x^*\| - \gamma_k \|R(x^k)\|)_+$$

$$\geq (\|x^k - x^*\| - \gamma_k L\|x^k - x^*\|)_+ ,$$

$$\mathbf{E}(\|x^{k+1} - x^*\|^2 \mid x^k) \geq (1 - \gamma_k L)_+^2 \|x^k - x^*\|^2 + \gamma_k^2 \sigma^2 .$$

Then

$$\mathbf{E}\|x^{k+1} - x^*\|^2 \geq (1 - \gamma_k L)_+^2 \mathbf{E}\|x^k - x^*\|^2 + \gamma_k^2 \sigma^2 .$$

The piecewise-quadratic function on the right attains a minimum with respect to γ_k for

$$\gamma_k^* = L\mathbf{E}\|x^k - x^*\|^2/(L^2\mathbf{E}\|x^k - x^*\|^2 + \sigma^2) ,$$

from which we obtain

$$\mathbf{E}\|x^{k+1} - x^*\|^2 \geq (1 - \gamma_k^* L)_+^2 \mathbf{E}\|x^k - x^*\|^2 + (\gamma_k^*)^2 \sigma^2$$

$$= \sigma^2 \mathbf{E}\|x^k - x^*\|^2/(L^2\mathbf{E}\|x^k - x^*\|^2 + \sigma^2) ,$$

or, denoting

$$u_k = 1/(\mathbf{E}\|x^k - x^*\|^2) , \quad u_{k+1} = L^2/\sigma^2 + u_k .$$

Thus, $u_k \leq u_0 + kL^2/\sigma^2$, i.e.,

$$\mathbf{E}\|x^k - x^*\|^2 \geq [1/\mathbf{E}\|x^0 - x^*\|^2 + kL^2/\sigma^2]^{-1} . \quad \square$$

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 $f(x)$:

$$x^{k+1} = x^k - \gamma_k s^k , \quad s^k = \nabla f(x^k) + \xi^k \tag{5}$$

under absolute random noise:

$$\mathbf{E}\xi^k = 0 , \quad \mathbf{E}\|\xi^k\|^2 \geq \sigma^2 \tag{6}$$

(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 γ_k, for method (5) one has the estimate

$$\mathbf{E}\|x^k - x^*\|^2 \geq (1/\mathbf{E}\|x^0 - x^*\|^2 + kL^2/\sigma^2)^{-1} . \tag{7}$$

Differently put, no variant of the gradient method under absolute random noises can converge faster than $O(1/k)$ (more precisely, $\mathbf{E}\|x^k - x^*\|^2 \geq \sigma^2/(L^2k) + o(1/k)$). 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 ξ^k. In this case Newton's method (modified by introduction of a parameter defining the step size) acquires the form

$$x^{k+1} = x^k - \gamma_k[\nabla^2 f(x^k)]^{-1}(\nabla f(x^k) + \xi^k) . \tag{8}$$

We shall assume the noise ξ^k is independent and

$$\mathbf{E}\xi^k = 0 , \quad \mathbf{E}\|\xi^k\|^2 \geq \sigma^2 . \tag{9}$$

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

$$\mathbf{E}\|\xi^k\|^2 \geq \tau\|x^k - x^*\|^2 . \tag{10}$$

Then for any method (1),

$$\mathbf{E}\|x^k - x^*\|^2 \geq \mathbf{E}\|x^0 - x^*\|^2 q^k , \quad q = \tau/(L^2 + \tau) . \ \square \tag{11}$$

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 γ_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:

$$f(x) = (Ax, x)/2 - (b, x), \quad \ell I \leq A \leq LI, \ \ell > 0 . \tag{12}$$

Consider the method

$$x^{k+1} = x^k - (\gamma_k/2(\alpha))(f(x^k + \alpha h^k) - f(x^k - ah^k))h^k , \tag{13}$$

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 γ_k may be chosen, the method of random search cannot converge more rapidly than the geomertic progression with ratio

$$q = (n-1)\ell^2/(L^2 + (n-1)\ell^2). \tag{14}$$

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

$$x^{k+1} = x^k - \Gamma_k(A(x^k - x^*) + \xi^k), \tag{15}$$

where ξ^k are independent, x^0 is a random vector, A^{-1} exists and

$$\mathbf{E}\xi^k = 0, \quad \mathbf{E}\xi^k(\xi^k)^T \geq B > 0, \quad \mathbf{E}(x^0 - x^*)(x^0 - x^*)^T > 0, \tag{16}$$

and Γ_k are deterministic $n \times n$ matrices.

THEOREM 3. In method (15) for any Γ_k one has the estimate

$$\begin{aligned}\mathbf{E}(x^k - x^*)(x^k - x^*)^T &\geq [(\mathbf{E}(x^0 - x^*)(x^0 - x^*)^T)^{-1} + kA^T B^{-1} A]^{-1} \\ &= \frac{1}{k} A^{-1} B (A^T)^{-1} + o\left(\frac{1}{k}\right). \ \square\end{aligned} \tag{17}$$

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:

$$x^{k+1} = x^k - \Gamma_k(\nabla f(x^k) + \xi^k), \quad \mathbf{E}\xi^k = 0, \ \mathbf{E}\xi^k(\xi^k)^T = \sigma^2 I. \tag{18}$$

Applying Theorem 3, we obtain for any Γ_k that

$$\mathbf{E}(x^k - x^*)(x^k - x^*)^T \geq \left(U_0^{-1} + \frac{k}{\sigma^2} A^2\right)^{-1} = \frac{\sigma^2}{k} A^{-1} + o\left(\frac{1}{k}\right), \tag{19}$$

$$U_0 = \mathbf{E}(x^0 - x^*)(x^0 - x^*)^T,$$

$$\|\mathbf{E}(x^k - x^*)(x^k - x^*)^T\| \geq \frac{\sigma^2}{k\ell^2} + o\left(\frac{1}{k}\right), \tag{20}$$

where equality obtains in (19), (20) (see Exercise 2) for

$$\Gamma_k = (kA + \sigma^2 A^{-1} U_0^{-1})^{-1} = k^{-1} A^{-1} + o(1/k). \tag{21}$$

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

1. Let h be a random vector uniformly distributed on the unit sphere in $\mathbf{R}^n$. Show that $\mathbf{E}hh^T = n^{-1}I$, and if a is an arbitrary vector, then $\mathbf{E}\|hh^T a - n^{-1}a\|^2 = (n^{-1} - n^{-2})\|a\|^2$.

2. Show that if $\mathbf{E}\xi^k(\xi^k)^T \vdash B$, then equality in (17) becomes equality for Γ_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

$$f(x) = (Ax, x)/2 - (b, x), \quad A > 0, \tag{22}$$

and its gradient be calculated with a random noise ξ. Suppose that the noise ξ 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, ..., x^k. Finally, let the matrices A and A^{-1} be known. Then $x^i - x^* = A^{-1}r^i - A^{-1}\xi^i$, $i = 1, \ldots, k$. 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 η^i are independent and identically distributed (since the ξ^i are). Thus the problem has been reduced to the following: Given the vectors $z^i = x^* + \eta^i$, where the η^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 η^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,

$$J = \int \frac{\nabla p_\eta(z) \nabla^T p_\eta(z)}{p_\eta(z)} dz, \quad 0 < J < \infty, \tag{23}$$

then for any unbiased estimate $\hat{x}^k$ of the vector x^* from the measurements z^i, $i = 1, \ldots, k$, the inequality holds:

$$\mathbf{E}(\hat{x}^k - x^*)(\hat{x}^k - x^*)^T \geq k^{-1}J^{-1}. \tag{24}$$

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 ξ^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 $r^i = \nabla f(x^i) + \xi^i$, $i = 1, \ldots, k$, at k points, the inequality holds:

$$\mathbf{E}(\hat{x}^k - x^*)(\hat{x}^k - x^*)^T \geq k^{-1} A^{-1} J^{-1} A^{-1} \, . \, \square \tag{25}$$

It is important that the measurement points $x^1, \ldots, x^k$ 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

$$x^{k+1} = x^k - \gamma_k H(\nabla f(x^k) + \xi^k) \, , \tag{26}$$

where $H > 0$ is some matrix, then the asymptotically optimal choice of γ_k and H are the following:

$$\gamma_k = 1/k \, , \quad H = A^{-1} \, , \tag{27}$$

and here

$$\mathbf{E}(x^k - x^*)(x^k - x^*)^T \leq k^{-1} A^{-1} B A^{-1} + o(k^{-1}) \, , \quad B = \mathbf{E}\xi\xi^T \, . \tag{28}$$

Noting Exercise 3, we obtain that if ξ^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

$$x^{k+1} = x^k - \gamma_k \phi(\nabla f(x^k) + \xi^k) \, , \tag{29}$$

where the function ϕ: $\mathbf{R}^n \to \mathbf{R}^n$ and γ_k are chosen as follows:

$$\phi(z) = -A^{-1} J^{-1} \nabla \log p(z) \, , \quad \gamma_k = 1/k \, . \tag{30}$$

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 ($k \ll \sigma^2 A^{-2} U_0^{-1}$) Γ_k is roughly constant: $\Gamma_k \approx \sigma^{-2} U_0 A$, while for large k, Γ_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 η 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 ξ and η be related by the equality $\eta = B\xi$, where B is some matrix. Prove that for the corresponding information matrices one has $J_\eta = BJ_\xi B^T$.

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

$$\begin{aligned} &x, y \in Q, \quad 0 \le \lambda \le 1, \\ &\lambda x + (1-\lambda)y \in Q. \end{aligned} \tag{1}$$

By induction, Q contains also any convex combination of points, i.e.,

$$\begin{aligned} &x^i \in Q, \quad \sum_{i=1}^{m} \lambda_i = 1, \quad \lambda_i \ge 0, \\ &i = 1, \ldots, m \Rightarrow \sum_{i=1}^{m} \lambda_i x^i \in Q. \end{aligned} \tag{2}$$

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).

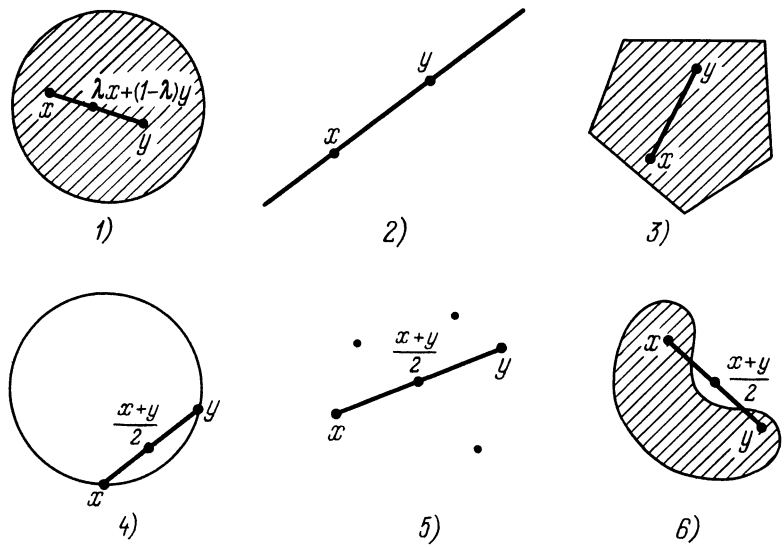


Fig. 10 Examples of convex (1-3) and nonconvex (4-6) sets.

For a convex function $f(x)$ the set $Q = \{x: f(x) \le \alpha\}$ is obviously convex for any α. The converse is not generally true: the function $f(x) = \sqrt{\|x\|}$ is not convex but the sets $\{x: f(x) \le \alpha\}$ are convex (such functions are called *quasiconvex*).

If a set Q is nonconvex, it can be "convexified." By the *convex hull* 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 nonempty 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.,

$$\text{Conv } Q = \left\{x = \sum_{i=1}^{m} \lambda_i x^i;\ x^i \in Q, \lambda_i \ge 0,\ i = 1,\ldots,m,\ \sum_{i=1}^{m} \lambda_i = 1\right\}. \quad (3)$$

LEMMA 1 (Caratheodory). For $Q \subset \mathbf{R}^n$ in (3) one can take $m \le n+1$. □

For a closed set Q the set Conv Q is not necessarily closed (for example, for

$$Q = \{x \in \mathbf{R}^2:\ x_2 = x_1^{1/2},\ x_1 \ge 0\}$$

one has

$$\text{Conv } Q = \{x \in \mathbf{R}^2:\ 0 < x_2 = x_1^{1/2},\ x_1 \ge 0\} \cup \{0, 0\}.$$

LEMMA 2. If Q is closed and bounded, so is Conv Q. □

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.,

$$P_Q(x) = \operatorname*{argmin}_{y \in Q} \|x - y\|. \quad (4)$$

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 $P_Q(x) = \operatorname*{argmin}_{y \in Q} \phi(y)$, $\phi(y) = \|x - y\|^2$ 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):

$$(x - P_Q(x), y - P_Q(x)) \le 0 \qquad \text{for all } y \in Q, \quad (5)$$

$$\|P_Q(x) - P_Q(y)\| \le \|x - y\| \qquad \text{for any } x, y. \quad (6)$$

Exercises

1. 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 $\alpha \in \mathbf{R}^1$, 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.
2. 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 $\nabla\phi(x) = x - P_Q(x)$.
3. 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 $0 \le \lambda < 1$ 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 $\mathbf{R}^n$ are called *separable* if there is a hyperplane separating them (Fig. 12), or, in other words, if there is a number α and a vector $a \in \mathbf{R}^n$, $a \neq 0$, such that $(a, x) \ge \alpha$ for all $x \in Q_1$ and $(a, x) \le \alpha$ for all $x \in Q_2$. These sets are strictly separable if there are $a \in \mathbf{R}^n$ and $\alpha_1 > \alpha_2$ such that $(a, x) \ge \alpha_1$ for $x \in Q_1$ and $(a, x) \le \alpha_2$ for $x \in Q_2$.

THEOREM 1 (separation theorem). Let Q_1, Q_2 be convex disjoint sets in $\mathbf{R}^n$ where Q_2 is bounded. Then Q_1 and Q_2 are strictly separable.

PROOF. The function

$$\rho_1(x) = P_{Q_1}(x) = \|x - P_{Q_1}(x)\|,$$

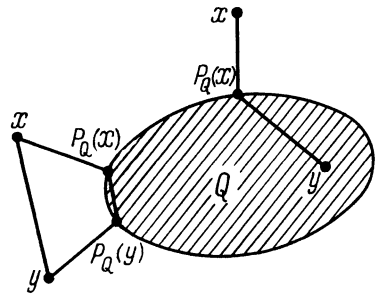


Fig. 11 Projections and its properties.

according to Exercise 2 above, is continuous. Hence, on the closed and bounded set Q_2 it attains a minimum. Let

$$a_1 = P_{Q_1}(a_2), \quad a_2 = \operatorname*{argmin}_{x \in Q_2} \rho_1(x).$$

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

$$(a_1 - a_2, x) \geq (a_1 - a_2, a_1) = \alpha_1 \quad \text{for } x \in Q_1,$$

$$(a_1 - a_2, x) \geq (a_1 - a_2, a_2) = \alpha_2 \quad \text{for } x \in Q_2,$$

$$\alpha_1 - \alpha_2 = \|a_1 - a_2\|^2 > 0.$$

Thus, Q_1 and Q_2 are strictly separable. □

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.

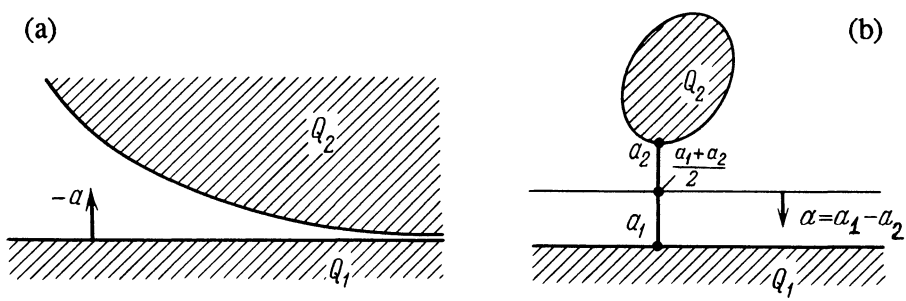


Fig. 12 Separation theorem: (a) separable sets; (b) strictly separable sets.

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).

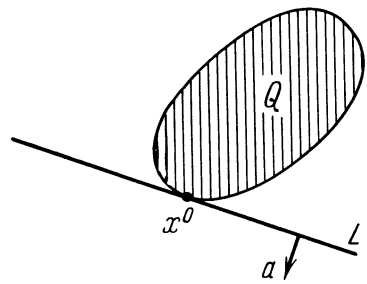


Fig. 13 The supporting hyperplane.

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. □

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 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

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \tag{7}$$

(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

$$a^1 = x + \delta e_1, \quad a^2 = x - \delta e_1, \quad \ldots, \quad a^{2n-1} = x + \delta e_n, \quad a^{2n} = \delta e_n ,$$

where $e_1, \ldots, e_n$ 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:

$$Q(\delta) = \left\{ \sum_{i=1}^{2n} \mu_i a^i, \ \mu_i \geq 0, \ \sum_{i=1}^{2n} \mu_i = 1 \right\} = \left\{ x + \delta \sum_{i=1}^{2n} \gamma_i e_i, \ |\gamma_i| \leq 1 \right\} .$$

Let us prove that

$$\sup_{y \in Q(\delta)} |f(y) - f(x)| \leq \Delta(\delta) .$$

Indeed, let

$$y = \sum_{i=1}^{2n} \mu_i a^i , \qquad \mu_i \geq 0 , \qquad \sum_{i=1}^{2n} \mu_i = 1 .$$

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 $\varepsilon > 0$,

$$\phi(\tau + \varepsilon) = \phi((1 - \varepsilon)\tau + \varepsilon(\tau + 1)) \leq (1 - \varepsilon)\phi(\tau) + \varepsilon\phi(\tau + 1)$$

$$= \phi(\tau) + \varepsilon(\phi(\tau + 1) - \phi(\tau)) ;$$

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 $\phi(\tau + \varepsilon) \to \phi(\tau)$ as $\varepsilon \to +0$. 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

$$\phi(\tau) - \phi(\tau - 1) \leq \phi'(\tau; 1) \leq \phi(\tau + 1) - \phi(\tau) . \tag{8}$$

By continuity,

$$\Delta_i(\delta) = |f(a^i) - f(x)| = |f(x \pm \delta e_i) - f(x)|$$

tends to 0 as $\delta \to 0$. Hence also $\Delta(\delta) = \max_i \Delta_i(\delta)$ tends to 0 as $\delta \to 0$. Therefore

$$\sup_{y \in Q(\delta)} |f(y) - f(x)| \to 0 \quad \text{as } \delta \to 0,$$

what was to be proved. □

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 $X^* = \operatorname{argmin}_{x \in \mathbf{R}^n} f(x)$ is convex and closed. □

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:

$$f'(x; y) = \lim_{\alpha \to +0} \frac{f(x+\alpha y) - f(x)}{\alpha} \leq f(x + y) - f(x)$$

$$\leq \max_{\|z\| = \|y\|} (f(x + z) - f(x)) . \quad \square \tag{9}$$

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 $\mathbf{R}^n$. A vector $a \in \mathbf{R}^n$ for which

$$f(x + y) \geq f(x) + (a, y) \tag{10}$$

for all $y \in \mathbf{R}^n$ 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.

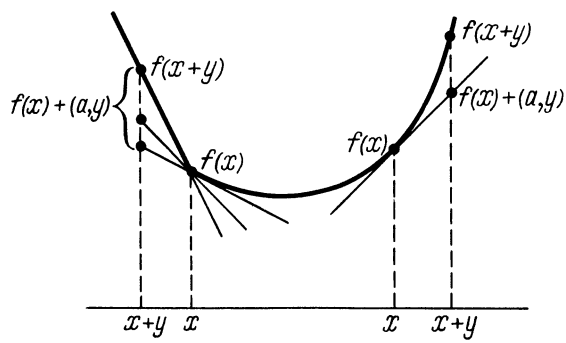


Fig. 14 The subgradient of a convex function.

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 $\partial f(x) - \nabla f(x) = 0$. □

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. □

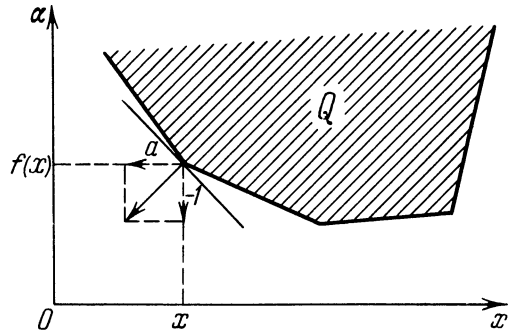


Fig. 15 Re: proof of the existence of the subgradient.

For nonsmooth functions we have the inequality similar to (29) in Section 1.1: for any x, y

$$(\partial f(x) - \partial f(y), x - y) \geq 0, \tag{11}$$

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

$$f'(x; y) = \max_{a \in \partial f(x)} (a, y) . \tag{12}$$

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. □

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 $\{x: f(x) \leq \alpha\}$. □

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. □

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 α_1, $\alpha_2 \geq 0$, then

$$\partial f(x) = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x) . \tag{13}$$

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)

$$\max_{a \in \partial f(x)} (a, y) = \max_{b \in \alpha_1 \partial f_1(x)} (b, y) + \max_{c \in \alpha_2 \partial f_2(x)} (c, y)$$

$$= \max_{a \in \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x)} (a, y) .$$

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 $y \in \mathbf{R}^n$, $\max_{a \in A} (a, y) = \max_{b \in B} (b, y)$ 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. □

Clearly, (13) extends to a sum of several convex functions:

$$\partial \left(\sum_{i=1}^{m} \alpha_i f_i(x) \right) = \sum_{i=1}^{m} \alpha_i \partial f_i(x) , \quad \alpha_i \geq 0 . \tag{14}$$

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

$$\partial f(x) = \operatorname{Conv} \bigcup_{i \in I(x)} \partial f_i(x) , \quad I(x) = \{i: f_i(x) = f(x)\} .$$

PROOF. By Lemmas 6, 2 the set

$$A = \operatorname{Conv} \bigcup_{i \in I(x)} \partial f_i(x)$$

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 Conv Q,

$$\max_{i \in I(x)} f_i'(x; y) = \max_{\lambda_i \geq 0, \sum_{i \in I(x)} \lambda_i = 1} \sum_{i \in I(x)} \lambda_i f_i'(x; y) = \max_{a \in A} (a, y) .$$

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) $A = \partial f(x)$. □

LEMMA 12. Let A be a an $m \times n$ matrix, let $\phi(y)$ be a convex function on $\mathbf{R}^m$, and let $f(x) = \phi(Ax)$, $x \in \mathbf{R}^n$. Then

$$\partial f(x) = A^T \partial \phi(Ax) . \quad \square \tag{16}$$

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) $f(x) = \max_{1 \leq i \leq n} ((a^i, x) - b_i)$.

ANSWERS: (a) $\partial f(x) = \begin{cases} \dfrac{x}{\|x\|}, & x \neq 0, \\ a, & \|a\| \leq 1, \ x = 0; \end{cases}$

(b) $\partial f(x) = \sum_{i=1}^{k} \operatorname{sign} ((a^i, x) - b_i) a^i$;

(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 ε-subgradient

The notion of a subgradient can be extended as follows. A vector $a \in \mathbf{R}^n$ is called the *ε-subgradient* of the convex function $f(x)$ at a point x if

$$f(x + y) \geq f(x) + (a, y) - \varepsilon \tag{17}$$

for all $y \in \mathbf{R}^n$. Here $\varepsilon \geq 0$ is a fixed number. The set of ε-subgradients as well as an arbitrary ε-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 ε-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 ε-subgradient for $\varepsilon > 0$ is not determined by local properties of $f(x)$. Clearly, the ε-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 ε, x.

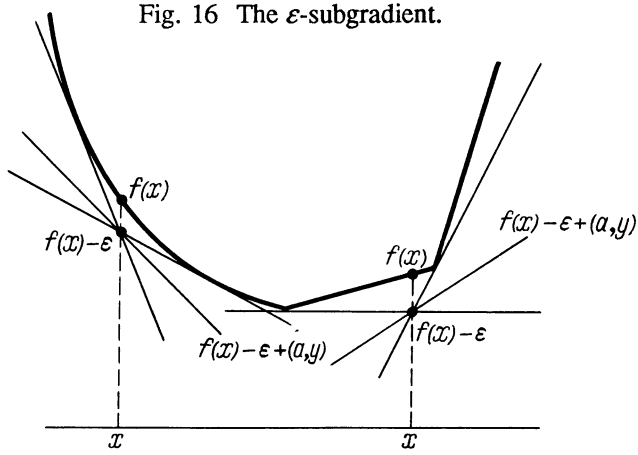
Rules for calculating the ε-subgradients are not as simple as for subgradients. We give here one important particular case where finding a ε-subgradient requires less calculations than the subgradient. Let

$$f(x) = \max_{y \in Q} \phi(x, y), \tag{18}$$

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 $\bar{y} = \bar{y}(x)$ be any point in Q such that

$$\phi(x, \bar{y}) \geq f(x) - \varepsilon. \tag{19}$$

Fig. 16 The ε-subgradient.



In other words, $\bar{y}$ is a arbitrary point at which the maximum of $\phi(x, y)$ in $y \in Q$ is attained aproximately (to within ε).

LEMMA 13.

$$\partial_x \phi(x, \bar{y}) \subset \partial_\varepsilon f(x) . \tag{20}$$

PROOF. For any z, by the definition of the subgradient and by (18), (19) we have

$$f(x+z) = \max_{y \in Q} \phi(x+z, y) \geq \phi(x+z, \bar{y}) \geq \phi(x, \bar{y}) + (\partial_x \phi(x, \bar{y}), z)$$

$$\geq f(x) + (\partial_x \phi(x, \bar{y}), z) - \varepsilon . \quad \square$$

Thus, to find one of the ε-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 ϕ, whereas the calculation of the subgradient of $f(x)$ requires that ϕ be maximized exactly in y.

5.2 EXTREMUM CONDITIONS, EXISTENCE, UNIQUENESS, AND STABILITY OF A SOLUTION

To analyze the problem

$$\min f(x) , \quad x \in \mathbf{R}^n , \tag{1}$$

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

$$0 \in \partial f(x^*) \tag{2}$$

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). $\square$

Of course, there may also be nonzero subgradients at a minimum point (e.g., for $f(x) = \|x\|$: $\partial f(0) = \{a: \|a\| \leq 1\}$, 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 ε-subgradient, we can formulate the necessary and sufficient conditions for the point x_ε to be an approximate solution of problem (1).

THEOREM 2. The condition

$$0 \in \partial_\varepsilon f(x_\varepsilon) \tag{3}$$

holds iff

$$f(x_\varepsilon) \leq \inf_x f(x) + \varepsilon . \quad \square$$

Exercises

1. Check the validity of the following extremum conditions:

(a) $f(x) = \sum_{i=1}^m \alpha_i \|x - a^i\|$, $\alpha_i > 0$, $x, a^i \in \mathbf{R}^n$. Then $\nabla f(x^*) = 0$, if $x^* \neq a^i$, and $\alpha_i \geq \|\sum_{j \neq i} \alpha_j (a^i - a^j)/\|a^i - a^j\|\|$, 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 $\sum_{i \in I^*} \lambda_i^* a^i + \sum_{i \in I_+} a^i - \sum_{i \in I_-} a^i = 0$, $I_+ = \{i: (a^i, x^*) \not< b_i\}$, $I_- = \{i: (a^i, x^*) \not> b_i\}$.

(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_ε 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 α. 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). □

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 $x \neq y$, $0 < \lambda < 1$,

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y) . \tag{4}$$

THEOREM 4. A minimum point of a strictly convex function is unique. The proof is obvious. □

Exercises

3. Prove that for a strictly convex function the following inequality holds for all $y \neq 0$:

$$f(x + y) > f(x) + (\partial f(x), y) . \tag{5}$$

4. Show that the function $f(x) = \|x\|$ is not strictly convex, whereas the function $\Sigma_{i=1}^m \alpha_i \|x-a^i\|$, $\alpha_i > 0$, is strictly convex provided the points a^i are not collinear.

5. Prove that the function

$$f(x) = \sum_{i=1}^{m} \alpha_i \|x-a^i\|, \qquad \alpha_i > 0 ,$$

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^*. □

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 $Q_\alpha = \{x: f(x) \le \alpha\}$ is bounded and nonempty for some α for the convex function $f(x)$, then Q_α is bounded for all α. □

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

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y) - \ell\lambda(1-\lambda)\|x - y\|^2/2 \qquad (6)$$

for any x, y and $0 \le \lambda \le 1$. Such functions have the following properties.

LEMMA 2. For a strongly convex function $f(x)$ we have

$$f(y) \ge f(x) + (\partial f(x), y-x) + \ell\|y - x\|^2/2 \qquad (7)$$

for all x, y, the $f(x)$ attaining a unique minimum x^* and for all x

$$f(x) \ge f(x^*) + \ell\|x - x^*\|^2/2 . \qquad (8)$$

PROOF. By the definition of a subgradient we have

$$f(\lambda x + (1-\lambda)y) = f(x + (1-\lambda)(y - x)) \ge f(x) + (1-\lambda)(\partial f(x), y-x) .$$

Substituting this inequality into (6) and cancelling out the term $1 - \lambda$ yield

$$f(y) \ge f(x) + (\partial f(x), y - x) + \ell\lambda\|x - y\|^2/2 .$$

This holds for all $\lambda < 1$; passing to the limit as $\lambda \to 1$, we obtain (7). It follows from (7) that $Q = \{y: f(y) \le f(x)\}$ is bounded, and Theorems 3 and 4 imply the existence and uniqueness of x^*. Using Theorem 1 together with (7), we arrive at (8). □

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)

$$f(x) \ge f(x^*) + \alpha\|x - x^*\| , \quad \alpha > 0 . \qquad (9)$$

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 $\partial f(x^*)$. □

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 $\mathbf{R}^n$, 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 ε 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 $\phi_\varepsilon(x)$. □

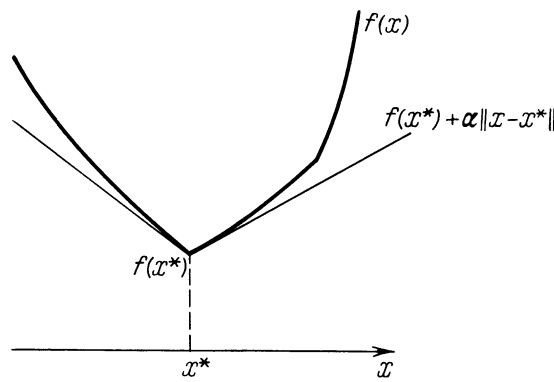


Fig. 17 The sharp minimum.

Exercises

6. Prove the following generalization of Theorem 6. Let $f(x)$ be a convex function and let $X^* = \operatorname{Argmin}_{x \in \mathbf{R}^n} f(x) \neq \emptyset$. Furthermore, let

$$f(x) \geq f^* + \alpha\rho(x, X^*), \qquad \alpha > 0,$$

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 $X_g^* = \operatorname{Argmin}_{x \in X^*} g(x)$ be nonempty and bounded. Then

$$X_g^* = \operatorname{Argmin}_{x \in \mathbf{R}^n} [f(x) + \varepsilon g(x)]$$

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 μ 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 the vector $s = s(x) \in \mathbf{R}^n$, $\|s\| = 1$, 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:

$$s(x) = \operatorname*{argmin}_{\|y\| = 1} f'(x; y). \tag{1}$$

By formula (12) of Section 5.1, for a convex function the direction of steepest descent exists and is defined by

$$s = -P_{\partial f(x)}(0)/\|P_{\partial f(x)}(0)\|, \tag{2}$$

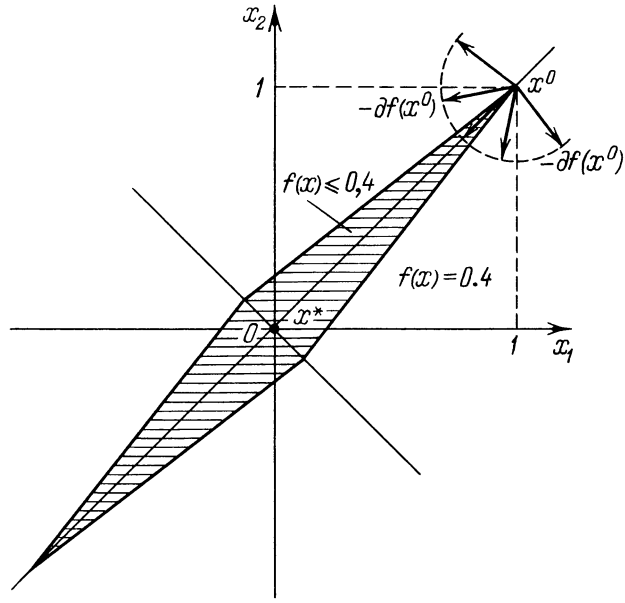


Fig. 18 Difficulties in minimizing a nonsmooth function.

i.e., s is the subgradient with minimal norm. However, it is possible to construct a convex function for which the steepest descent method

$$x^{k+1} = x^k + \gamma_k s(x^k), \quad \gamma_k = \operatorname*{argmin}_{\gamma > 0} f(x^k + \gamma s(x^k))$$

"jams" without reaching the minimim 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:

$$x^{k+1} = x^k - \gamma_k \partial f(x^k). \tag{3}$$

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 γ_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 γ_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:

$$x^{k+1} = x^k - \gamma_k \frac{\partial f(x^k)}{\|\partial f(x^k)\|}, \quad \gamma_k \to 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty. \tag{4}$$

In other words, the step of fixed size γ_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 γ_k satisfying conditions (4) are given by

$$\gamma_k = \frac{\gamma}{k+c}, \quad \gamma_k = \frac{\gamma}{k^\rho}, \quad 0 < \rho \leq 1, \quad \gamma_k = \frac{\gamma}{k \ln k}. \tag{5}$$

The assertion on convergence in Theorem 1 (and in many cases in the sequel) concerns the quantity

$$\phi_k = \min_{0 \leq i \leq k} f(x^i) \tag{6}$$

being the *record value* of $f(x)$ over k iterations.

THEOREM 1. In method (4) for convex $f(x)$: $\phi_k \to f^* = \inf_{x \in \mathbf{R}^n} f(x)$.

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 $\bar{x}$ such that $f(\bar{x}) < \tilde{f}$. 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 $\|x - \bar{x}\| \leq \rho$. In particular, for

$x_\rho = \bar{x} + \rho \partial f(x^k)/\|\partial f(x^k)\|$ we have $f(x_\rho) \le \tilde{f}$. On the other hand,

$$\begin{aligned} f(x_\rho) &\ge f(x^k) + (\partial f(x^k), x_\rho - x^k) \ge \\ &\ge \tilde{f} + (\partial f(x^k), \bar{x} - x^k) + (\partial f(x^k), x_\rho - \bar{x}) \\ &= \tilde{f} + (\partial f(x^k), \bar{x} - x^k) + \rho \|\partial f(x^k)\| , \end{aligned}$$

i.e.,

$$(\partial f(x^k), x^k - \bar{x})/\|\partial f(x^k)\| \ge \rho .$$

Let us now estimate the distance to $\bar{x}$ in the iterations:

$$\begin{aligned} \|x^{k+1} - \bar{x}\|^2 &= \|x^k - \bar{x}\|^2 - 2\gamma_k \left(\frac{\partial f(x^k)}{\|\partial f(x^k)\|}, x^k - \bar{x} \right) + \gamma_k^2 \\ &\le \|x^k - \bar{x}\|^2 - 2\gamma_k \rho + \gamma_k^2 . \end{aligned}$$

Since $\gamma_k \to 0$, we can find a k_0 such that $\gamma_k \le \rho$ for $k \ge k_0$. Hence for $k \ge k_0$, we have

$$\|x^{k+1} - \bar{x}\|^2 \le \|x^k - \bar{x}\|^2 - \gamma_k \rho .$$

Summing these inequalities over k, we obtain $\rho \Sigma_{k=k_0}^{\infty} \gamma_k \le \|x^{k_0} - \bar{x}\|^2$, which contradicts the condition $\Sigma_{k=0}^{\infty} \gamma_k = \infty$. Thus, the inequality $f(x^k) \ge \tilde{f} > f^*$ is impossible for all k, which is equivalent to the condition $\phi_k \to f^*$. □

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 γ_k; this quantity decreases slowly since the condition $\Sigma_{k=0}^{\infty} \gamma_k = \infty$ 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 γ_k from the conditions $\gamma_k \to 0$, $\Sigma_{k=0}^{\infty} \gamma_k = \infty$ 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

$$(a^i, x) = b_i , \quad i = 1,...,n , \quad x \in \mathbf{R}^n ,$$

is reduced to a minimization of the function

$$f(x) = \sum_{i=1}^{n} |(a^i, x) - b_i|$$

or of the function

$$f(x) = \max_{1 \le i \le n} |(a^i, x) - b_i| ,$$

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:

$$x^{k+1} = x^k - \frac{f(x^k) - f^*}{\|\partial f(x^k)\|^2} \partial f(x^k) . \tag{7}$$

This choice of the step size is shown graphically in Figure 19.

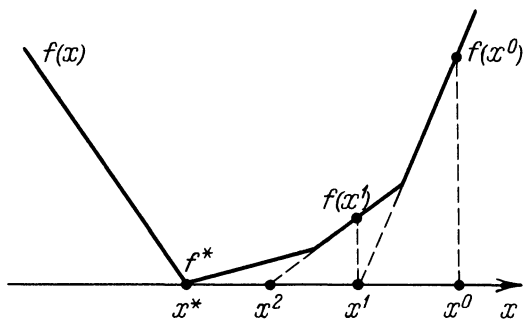


Fig. 19 The method for choosing a step size in the subgradient method.

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), $x^k \to x^* \in X^*$. The estimation of the convergence rate is as follows: for an arbitrary function f,

$$\lim_{k \to \infty} \sqrt{k} \, (f(x^k) - f^*) = 0 ; \tag{8}$$

for the function with a sharp minimum one can claim convergence with the rate of geometric progression.

PROOF. Let $\bar{x}$ be an arbitrary minimum point. Then

$$\|x^{k+1} - \bar{x}\|^2 = \|x^k - \bar{x}\|^2 - 2\frac{(\partial f(x^k),\ x^k - \bar{x})\,(f(x^k) - f^*)}{\|\partial f(x^k)\|^2} + \frac{(f(x^k) - f^*)^2}{\|\partial f(x^k)\|^2}$$

$$\leq \|x^k - \bar{x}\|^2 - \frac{(f(x^k) - f^*)^2}{\|\partial f(x^k)\|^2} \tag{9}$$

and $(f(x^k) - f^*)/\|\partial f(x^k)\| \to 0$. Since the sequence x^k is bounded: $\|x^k - \bar{x}\| \leq \|x^0 - \bar{x}\|$, then (Lemma 8 of Section 5.1) $\|\partial f(x^k)\| \leq c$. Hence $f(x^k) \to f^*$. Therefore, we can find a sequence $x^{k_i} \to x^*$, where x^* is a minimum point. If in the foregoing estimate we replace $\bar{x}$ by x^*, $\|x^k - x^*\|$ will monotonically decrease, whereas $\|x^{k_i} - x^*\| \to 0$. Thus $x^k \to x^*$.

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 $\Sigma_{k=0}^{\infty}(f(x^k) - f^*)^2 < \infty$. If we assume that $\varliminf_{k\to\infty} \sqrt{k}\,(f(x^k) - f^*) > 0$, then $f(x^k) - f^* > a/\sqrt{k}$ for sufficiently large k, which contradicts the condition $\Sigma_{k=0}^{\infty}(f(x^k) - f^*)^2 < \infty$. Thus, $\varliminf_{k\to\infty} \sqrt{k}\,(f(x^k) - f^*) = 0$.

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. □

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

$$x^{k+1} = x^k - \frac{f(x^*) - \bar{f}}{\|\partial f(x^k)\|^2}\,\partial f(x^k), \tag{10}$$

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 $\rho(x^k, X^*) \to 0$.

(b) If $\Sigma_{k=0}^{\infty} \gamma_k^2 < \infty$, then $x^k \to x^* \in X^*$.

(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) $x^{k+1} = x^k - \gamma_k \partial f(x^k)$, $\gamma_k \to 0$, $\Sigma_{k=0}^{\infty} \gamma_k = \infty$;

(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 γ," i.e., there exists a function $\psi(\gamma) > 0$, $\psi(\gamma) \to 0$ as $\gamma \to 0$ such that

$$\lim_{k \to \infty} \phi_k \leq f^* + \psi(\gamma), \qquad \phi_k = \min_{1 \leq i \leq k} f(x^i).$$

(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 γ_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

$$\phi_{km} = c(\|x^m - x^*\|^2 + \Sigma_{i=m}^{\infty} \gamma_i^2)/\Sigma_{i=m}^{k} \gamma_i$$

and choose $m = k/2$ for even k.

5.3.3 The ε-subgradient Method

Examine whether it is possible to replace the subgradient by the ε-subgradient in methods of the form (3). It would be appropriate to do

so because in many problems it is easier to calculate the ε-subgradient than the gradient (see Lemma 13 of Section 5.1).

The most straightforward approach involves a substitution of an arbitrary ε-subgradient for the $\partial f(x)$ in method (4). However, if ε is fixed, then the new method may not converge: e.g., by Theorem 2 of Section 5.2 the ε-subgradient may vanish at any point at which the value of $f(x)$ differs from the optimum by less than ε—the method can therefore stop at such points. To make this method converge, we need to vary ε by letting it go to 0 in the iterative process. Then we obtain the following method:

$$x^{k+1} = x^k - \gamma_k \frac{\partial_{\varepsilon_k} f(x^k)}{\|\partial_{\varepsilon_k} f(x^k)\|}, \qquad \gamma_k \to 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty, \quad \varepsilon_k \to 0 \quad \text{as} \quad k \to \infty. \tag{11}$$

THEOREM 3. In method (11) for the convex function $f(x)$ we have

$$\phi_k = \min_{1 \le i \le k} f(x^k) \to \inf f(x).$$

Proof follows that of Theorem 1. □

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 $x^0, \ldots, x^k$ have been constructed and the subgradients $\partial f(x^0), \ldots, \partial f(x^k)$ 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

$$Q_k = \{x: (\partial f(x^i), x - x^i) \leq f^* - f(x^i), i = 0,\ldots,k\} , \qquad (1)$$

and for the unknown $f^* = f(x^*)$ it lies in the broader domain

$$Q_k = \{x: (\partial f(x^i), x - x^i) \leq 0, i = 0,\ldots,k\} . \qquad (2)$$

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

$$f^* = \min_{x \in \mathbf{R}^n} f(x) , \qquad \phi_k = \min_{0 \leq i \leq k} f(x^i) .$$

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 $\partial f(x^i)$, $i = 0, \ldots, k$, on the set Q_0. In other words, x^{k+1} is the solution of the linear programming problem:

$$\min z ,$$

$$f(x^i) + (\partial f(x^i), x - x^i) \leq z , \quad i = 0,...,k , \ x \in Q_0 . \tag{3}$$

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.

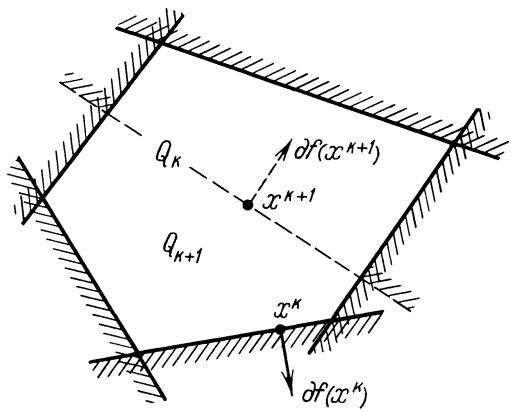


Fig. 20 A general scheme of the cutoff method.

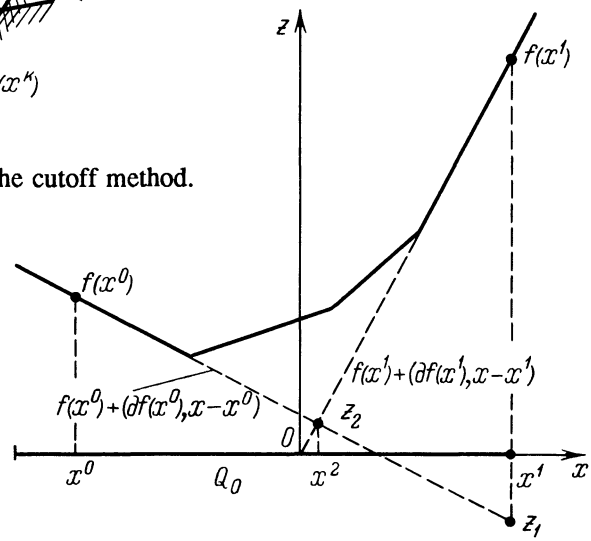


Fig. 21 The cutting hyperplane method.

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 $\phi_k \to f^*$.

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 $i = 0, \ldots, k$ and all $k \geq k_0$. This contradicts the compactness of Q_0. Hence

$$\lim_{k \to \infty} (f(x^k) - z_k) = 0 ,$$

and since

$$0 \leq f(x^k) - f^* \leq f(x^k) - z_k ,$$

then

$$\lim_{k\to\infty} f(x^k) = f^* . \ \square$$

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

$$\min p^{-1}x^p , \quad 0 \leq x \leq 1, \ x_0 = 0, \ x_1 = 1 .$$

Each auxiliary problem of (3) has a nonunique solution, the x_{k+1} being the largest solution. Then $x_{k+1} = x_k - p^{-1}x_k = qx_k$, $x_k = q^{k-1}$, $q = 1-p^{-1}$, 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

$$\max z ,$$

$$\left(\frac{\partial f(x^i)}{\|\partial f(x^i)\|}, x - x^*\right) + z \leq 0 , \quad i = 0,...,k, \ x \in Q_0 , \tag{4}$$

or, if f^* is known, of the problem

$$\max z ,$$

$$\left(\frac{\partial f(x^i)}{\|\partial f(x^i)\|}, x - x^*\right) + z \leq f^* - f(x^i) , \quad i = 0,...,k, \ x \in Q_0 . \tag{5}$$

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.

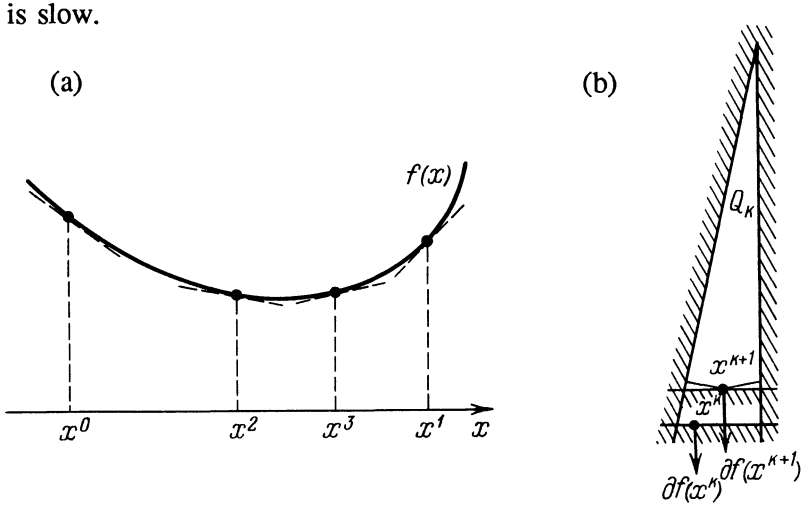


Fig. 22 The method of Chebyshev centers:
(a) one-dimensional case; (b) two-dimensional case.

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 $0,\ldots,k$, for instance, $I = \{0,\ldots,k\}$, 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

$$\min \|x - x^k\|^2 ,$$

$$f(x^i) + (\partial f(x^i), x - x^i) \leq f^* , \quad i \in I \tag{6}$$

(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)

$$x^{k+1} = x^k - \sum_{i \in I} \lambda_i^k \partial f(x^i) , \tag{7}$$

where λ_i^k is the solution of the problem

$$\min_{\substack{\lambda_i \geq 0 \\ i \in I}} \left[\left\| \sum_{i \in I} \lambda_i \partial f(x^i) \right\|^2 - 4 \sum_{i \in I} \lambda_i (\partial f(x^i), x^k - x^i) - 4 \sum_{i \in I} \lambda_i (f(x^i) - f^*) \right] . \tag{8}$$

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

$$Q_k = \{x \in Q_0 : (\partial f(x^i), x - x^i) \leq 0, \ i = 1,...,k\} , \tag{9}$$

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 $\mathbf{R}^n$, 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

$$\frac{v_i}{v} \leq 1 - \left(\frac{n}{1+n}\right)^n < 1 - \frac{1}{e}, \quad i = 1,2, \ v = v_1 + v_2 . \tag{10}$$

For points other than a the right side in (10) can only be greater. □

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.

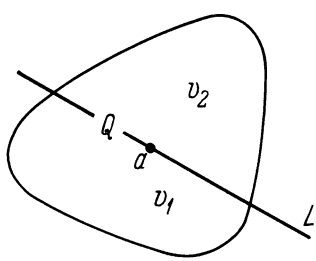


Fig. 23 The lemma on the center of gravity.

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

$$\phi_k - f^* \le cq^k ,$$

$$q = \left(1 - \left(\frac{n}{n+1}\right)^n\right)^{1/n} < \left(1 - \frac{1}{e}\right)^{1/n} = 1 - \frac{1}{ne} + o\left(\frac{1}{n}\right) ,$$

$$c = \max_{x \in Q_0} (f(x) - f^*) . \tag{11}$$

PROOF. By Lemma 1, the volume v_k of the polyhedron Q_k satisfies the inequality $v_{k+1} \le v_k\beta$, $\beta = 1 - (n/(n+1))^n$, i.e., $v_k \le v_0\beta^k$. 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 $v(S) = \alpha^n v_k \le \alpha^n v_0\beta^k = v_0$. 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, $1 \le i \le k$, such that $(\partial f(x^i), u - x^i) \ge 0$. Hence

$$f(u) \ge f(x^i) + (\partial f(x^i), u - x^i) \ge f(x^i) \ge \phi_k .$$

Using the convexity of $f(x)$, we obtain

$$\phi_k \le f(u) = f((1-\alpha^{-1})x^* + \alpha^{-1}z)$$

$$\le (1 - \alpha^{-1})f^* + \alpha^{-1}f(z) \le f^* + c/\alpha ,$$

$$c = \max_{x \in Q_0} (f(x) - f^*) ,$$

where $c < \infty$ since the $f(x)$ is continuous and the Q_0 is bounded. Thus

$$\phi_k - f^* \le c\alpha^{-1} = c\left(1 - \left(\frac{n}{n+1}\right)^n\right)^{k/n} . \quad \square$$

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 $\bar{x} = \alpha\bar{x}_1 + (1-\alpha)\bar{x}_2$, where $\bar{x}$, $\bar{x}_1$, $\bar{x}_2$ 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, $I = \{k,...,k-m\}$ 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). □

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 $x^0, \ldots, x^k, f(x^0), \ldots, f(x^k), \partial f(x^0), \ldots, f(x^k)$, 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 $f(x^k)$.

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 $\max_{1 \le i \le k} f_i(x)$, 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 $f(x^i)$ and $\partial f(x^i)$ 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 $\partial f(x^i)$, $i = 0,...,k$. It suffices to store at the kth 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*:

$$\begin{aligned} x^{k+1} &= x^k - \gamma_k H_k \partial f(x^k), \\ \gamma_k &= \frac{\rho}{n+1}\left[\frac{4}{\sqrt{n^2-1}}\right]^k (H_k \partial f(x^k), \partial f(x^k))^{-1/2}, \\ H_{k+1} &= H_k - \frac{2}{n+1}\frac{H_k \partial f(x^k)\, \partial f(x^k)^T H_k}{(H_k \partial f(x^k), \partial f(x^k))}, \quad H_0 = I, \end{aligned} \tag{12}$$

where ρ 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 $\mathbf{R}^n$; $n \geq 2$:

$$\begin{aligned} \phi_k - f^* &\leq cq^k, \quad c = \max_{\|x-x^0\| \leq \rho} (f(x) - f^*), \\ q &= n(n-1)^{-(n-1)/2n}\,(n+1)^{-(n+1)/2n}. \end{aligned} \tag{13}$$

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. □

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.

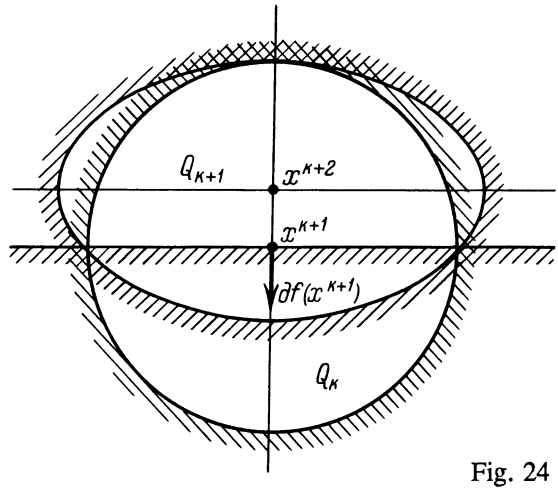


Fig. 24 The ellipsoid method.

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:

$$\begin{aligned} x^{k+1} &= x^k - \gamma_k H_k \partial f(x^k) \\ H_{k+1} &= H_k - \left(1 - \frac{1}{\alpha_k^2}\right) \frac{H_k s^k (s^k)^T H_k}{(H_k s^k, s^k)}, \qquad H_0 = I, \end{aligned} \tag{14}$$

where α_k is the space extension coefficient in the kth iteration, γ_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: $x^n = x^* = A^{-1}b$. □

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

$$N(f(x) - f^*) \leq (\partial f(x), x - x^*) \leq M(f(x) - f^*) . \tag{18}$$

These functions are referred to as *approximately homogeneous* (cf. (30) of Section 3.3). It can be proved for these functions that if

$$\alpha_k \equiv \alpha = (M + N)/(M - N) , \quad \lambda = 2MN/(M + N) , \tag{19}$$

the method (14), (16) converges with the rate of geometric progression with ratio $\alpha^{1/n}$:

$$\phi_k - f^* \leq c\sqrt{k}\,\alpha^{-k/n} . \tag{20}$$

Thus, the closer M is to N (i.e., the closer the function is to being homogeneous), the larger α 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 α 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

$$x^{k+1} = x^k - \gamma_k s^k, \quad s^k = \partial f(x^k) + r^k, \tag{1}$$

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 γ_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

$$\|r^k\| \leq \varepsilon, \tag{2}$$

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 ε 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 γ_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

$$\|x^k - x^*\| \leq \|x^0 - x^*\| q^k. \quad \square \tag{3}$$

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, α, $\varepsilon, \|x^0 - x^*\|$). Without this information, the incorrect choice of the γ_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:

$$\|r^k\| \leq \alpha \|\partial f(x^k)\| . \tag{4}$$

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

$$V(x) = \|x - x^*\|^2/2 \tag{5}$$

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 ϕ_k is the angle between the $\partial f(x^k)$ and $x^k - x^*$, $0 \leq \phi_k \leq \pi/2$. Hence, if

$$0 \leq \phi_k \leq \phi < \pi/2 , \qquad \alpha \leq \cos \phi , \tag{6}$$

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

$$x^{k+1} = x^k - \gamma_k H \partial f(x^k) , \tag{7}$$

$H > 0$ 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:

$$\mathbf{E} r^k = 0 , \qquad \mathbf{E}\|r^k\|^2 \leq \sigma^2 . \tag{8}$$

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

$$\sum_{k=0}^{\infty} \gamma_k = \infty, \quad \sum_{k=0}^{\infty} \gamma_k^2 < \infty. \tag{9}$$

Then in method (1)

$$\min_{0 \leq i \leq k} f(x^i) \to f(x^*) \text{ a.s. } \square$$

Thus, as in the smooth case, the method converges in additive random noise of any level if the γ_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 $\gamma_k \to 0$ instead of $\gamma_k \equiv \gamma$), whereas in the nonsmooth case the noise has little effect (noise, or no noise, one needs to take $\gamma_k \to 0$). 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 γ 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 ℓ, 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 $L = \max_{x \in S} \|\partial f(x)\|$, 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)):

$$a_0 = a\,, \quad b_0 = b\,, \quad \varepsilon_k = \alpha(b_k - a_k)/2\,, \quad 0 < \alpha < 1\,,$$

$$a_{k+1} = \begin{cases} a_k & \text{if } f((a_k+b_k)/2-\varepsilon_k) < f((a_k+b_k)/2+\varepsilon_k)\,, \\ (a_k+b_k)/2 - \varepsilon_k & \text{otherwise}\,, \end{cases}$$

$$b_{k+1} = \begin{cases} b_k & \text{if } f((a_k+b_k)/2-\varepsilon_k) > f((a_k+b_k)/2+\varepsilon_k)\,, \\ (a_k+b_k)/2 + \varepsilon_k & \text{otherwise}\,, \end{cases} \tag{1}$$

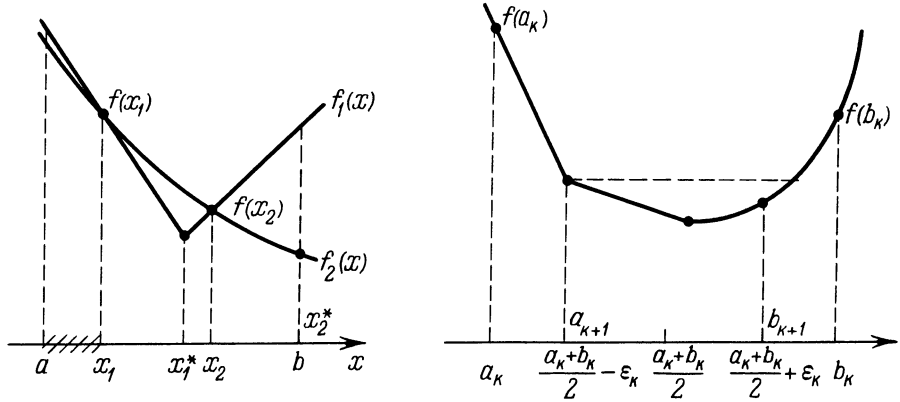
$$\left.\begin{aligned} a_{k+1} &= (a_k + b_k)/2 - \varepsilon_k\,, \\ b_{k+1} &= (a_k + b_k)/2 + \varepsilon_k\,, \end{aligned}\right\}$$

$$\text{if } f((a_k+b_k)/2 - \varepsilon_k) = f((a_k+b_k)/2 + \varepsilon_k)\,.$$

Obviously,

$$0 \le b_{k+1} - a_{k+1} \le (1 + \alpha)(b_k - a_k)/2\,,$$

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 α 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 α is chosen from the relation

$$(1 + \alpha)/2 = \beta, \quad \beta^2 = 1 - \beta, \quad \beta = (\sqrt{5} - 1)/2 \qquad (2)$$

(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 α 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

$$\begin{aligned} x^{k+1} &= x^k - \gamma_k s^k , \\ s^k &= \delta_k^{-1}[f(x^k + \alpha_k g^k + \delta_k h^k) - f(x^k + \alpha_k g^k)]h^k , \end{aligned} \tag{3}$$

where g^k, h^k are independent random vectors uniformly distributed on the cube $Q = \{x: |x_i| \le 1,\ i = 1,\ldots,n\}$, α_k, δ_k, γ_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

$$\mathbf{E}(s^k \mid x^k) = c\nabla f(x^k, \alpha_k) + \beta_k , \qquad \|\beta_k\| \le c_1 \delta_k/\alpha_k , \tag{4}$$

where $f(x, \alpha)$ is the *smoothed* function,

$$f(x, \alpha) = \frac{1}{(2\alpha)^n} \int_Q f(x + \alpha y)\, dy , \tag{5}$$

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 α_k, the size of the trial step δ_k and of the working step γ_k, one can get the method to converge. Thus, if

$$\begin{aligned} &\sum_{k=0}^{\infty} \gamma_k = \infty , \quad \sum_{k=0}^{\infty} \gamma_k^2 < \infty , \quad \gamma_k/\alpha_k \to 0 , \\ &\delta_k/\alpha_k \to 0 , \quad \alpha_k \to 0 , \quad |\alpha_k - \alpha_{k+1}|/\gamma_k \to 0 , \end{aligned} \tag{6}$$

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 $f(x)$:

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

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 $\nabla f(x^k) \to 0$ for (1). Thus, for convex functions a stronger result holds true.

THEOREM 1. Let $f(x)$ be a convex differentiable function in $\mathbf{R}^n$ whose gradient satisfies a Lipschitz condition with constant L, and the set of minimum points $X^* = \operatorname{Argmin}_{x \in \mathbf{R}^n} f(x)$ is nonempty. Then method (1) with $0 < \gamma < 2/L$ converges to some point $\bar{x} \in X^*$, $f(\bar{x}) = f^*$, with

$$f(x^k) - f^* = o(1/k) . \tag{2}$$

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

$$\begin{aligned} \|x^{k+1} - \hat{x}\|^2 &= \|x^k - \hat{x}\|^2 - 2\gamma(\nabla f(x^k), x^k - \hat{x}) + \gamma^2 \|\nabla f(x^k)\|^2 \\ &\leq \|x^k - \hat{x}\|^2 - \gamma(2/L - \gamma)\|\nabla f(x^k)\|^2 . \end{aligned} \tag{3}$$

Summing over k, we obtain that for $0 < \gamma < 2/L$

$$\sum_{k=0}^{\infty} \|\nabla f(x^k)\|^2 < \infty , \tag{4}$$

i.e., $\nabla f(x^k) \to 0$. 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 $x^{k_i} \to \bar{x}$. By the continuity of $\nabla f(x)$ we have here $\nabla f(\bar{x}) = 0$, i.e., $\bar{x} \in X^*$. Replacing the $\hat{x}$ with $\bar{x}$ in (3) yields $x^k \to \bar{x}$.

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

$$f(x): \ f(x^k) - f^* \leq (\nabla f(x^k), x^k - \bar{x}) \leq \|\nabla f(x^k)\| \ \|x^k - \bar{x}\| .$$

Hence for $u_k = f(x^k) - f^*$ we obtain $u_{k+1} \leq u_k - \alpha \|x^k - \bar{x}\|^{-2} u_k^2$, and applying Lemma 6 of Section 2.2 we have

$$u_k \leq \left(\frac{1}{u_0} + \alpha \sum_{i=0}^{k-1} \|x^i - \bar{x}\|^{-2}\right)^{-1},$$

$$ku_{k+1} \leq \left(\frac{1}{u_0 k} + \frac{\alpha}{k} \sum_{i=0}^{k-1} \|x^i - \bar{x}\|^{-2}\right)^{-1}.$$

Since according to what has been proved $\|x^i - \bar{x}\| \to 0$ as $i \to \infty$, we have:

$$\|x^i - \bar{x}\|^{-2} \to \infty \quad \text{and} \quad k^{-1} \sum_{i=0}^{k-1} \|x^i - \bar{x}\|^{-2} \to \infty \quad \text{as} \quad k \to \infty.$$

Therefore the right-hand side of the last inequality tends to zero as $k \to \infty$. This implies $u_k = o(1/k)$. □

We note that it is impossible to single out in advance the point $\bar{x} \in X^*$ to which x^k converges. For example, this point can vary for varied γ (for fixed x^0) and does not necessarily coincide with x^*, which is the point of X^* closest to x^0. However, $\bar{x}$ cannot lie too far away from x^*. Indeed, substituting x^* for $\hat{x}$ in (3), we obtain $\|x^k - x^*\| \leq \|x^0 - x^*\|$, i.e.,

$$\|\bar{x} - x^*\| \leq \|x^0 - x^*\| = \rho(x^0, X^*). \tag{5}$$

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 $\gamma > 0$, $\varepsilon_k > 0$, $\varepsilon_k \to 0$ 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 ε_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:

$$f(x) = (Ax, x)/2 - (b, x), \qquad A > 0. \tag{6}$$

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 ℓ 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

$$\|x^k - x^*\| \leq \|x^0 - x^*\| q^k, \qquad q = (L-\ell)/(L+\ell) = (\mu-1)/(\mu+1),$$

this estimate being unimprovable (see the examples after that theorem). Since

$$2(f(x^k) - f^*) = (A(x^k - x^*), x^k - x^*) \leq \|A\| \, \|x^k - x^*\|^2$$
$$\leq L\|x^0 - x^*\|^2 q^{2k},$$

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

$$f(x^k) - f^* \leq \frac{\|x^0 - x^*\|^2}{2\gamma(2k+1)} \left(1 - \frac{1}{2k+1}\right)^{2k} < \frac{\|x^0 - x^*\|^2}{4\gamma ek}. \qquad (7)$$

PROOF.

$$\begin{aligned} x^k - x^* &= (I - \gamma A)^k (x^0 - x^*), \\ 2(f(x^k) - f^*) &= (A(I - \gamma A)^{2k}(x^0 - x^*), x_0 - x^*) \\ &\leq \|x^0 - x^*\|^2 \|A(I - \gamma A)^{2k}\| \\ &\leq \|x^0 - x^*\|^2 \max_{\ell \leq \lambda \leq L} |\lambda(1 - \gamma\lambda)^{2k}| \\ &\leq \|x^0 - x^*\|^2 \max_{0 \leq \lambda \leq L} \phi(\lambda), \end{aligned}$$

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 $[0, L]$ can be obtained either for $\lambda = \lambda_2$, or for $\lambda = L$. Since

$$\phi(\lambda_2) = \frac{1}{\gamma(2k+1)} \left(1 - \frac{1}{2k+1}\right)^{2k} < \frac{1}{2\gamma\, ek}.$$

$$\phi(L) = L(1 - \gamma L)^{2k}, \quad \text{while } |1 - \gamma L| < 1,$$

then for sufficiently large k we have $\max_{0 \leq \lambda \leq L} \phi(\lambda) = \phi(\lambda_2)$, yielding (7). □

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 γ. 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:

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

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

$$f(x^k) - f^* \leq \frac{L\|x^0 - x^*\|^2}{2(2k + 1)^2} . \tag{8}$$

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 $\max_{0 \leq x \leq 1} |T_k(x)| = 1$. □

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:

$$\begin{aligned} x^k &= y^k - \gamma \nabla f(y^k) , & \gamma &= 1/L, \ y^1 = x^0 , \\ y^{k+1} &= x^k + \beta_k (x^k - x^{k-1}) , & \beta_k &= (\alpha_k - 1)/\alpha_{k+1} , \\ \alpha_{k+1} &= (\sqrt{4\alpha_k^2 + 1} + 1)/2 , & \alpha_1 &= \tfrac{1}{2} . \end{aligned} \tag{9}$$

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 $f'(x) = p|x|^{p-1}$ 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 $x_{k+1} = x_k - (p-1)^{-1} x_k = q x_k$, $q = (p-2)/(p-1) < 1$. 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

1. For a fixed number of steps k, find the best way of choosing the parameter γ in method (1) for minimizing (6), starting with the estimates derived in proving Theorem 2.

Hint: Take γ such as to minimize the $\max_{0 \leq \lambda \leq L} \lambda(1-\lambda)^{2k}$ for known k, L.

2. 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) γ_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 γ_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 $f(x) = \exp(-x^{-2})$, $x \in \mathbf{R}^1$, in a neighborhood of the minimum point $x^* = 0$.

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 ε 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 ε 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 minimim of the auxiliary problem is exact.

THEOREM 4. Let $f(x)$ be a convex continuous function in $\mathbf{R}^n$ having a nonempty set of minimum points X^*, and let $g(x)$ be a strongly convex continuous function. Let

$$x_\varepsilon = \operatorname*{argmin}_{x \in \mathbf{R}^n} \Phi_\varepsilon(x), \quad \Phi_\varepsilon = f(x) + \varepsilon g(x), \quad \varepsilon > 0. \tag{11}$$

Then $x_\varepsilon \to x^*$ as $\varepsilon \to +0$, where x^* is the minimum point of $f(x)$ for which $g(x)$ is minimal, i.e., $x^* = \operatorname*{argmin}_{x \in X^*} g(x)$.

PROOF. The function $f + \varepsilon g$ is strongly convex. Hence x_ε exists and is unique. Furthermore, from the definition of x_ε 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_ε is bounded. Take a subsequence x_{ε_i} converging to a point $\bar{x}$. Since $f(x)$ and $g(x)$ are continuous, then $\lim_{i \to \infty} g(x_{\varepsilon_i}) = g(\bar{x})$, i.e., $g(\bar{x}) \leq g(x^*)$, 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 $f(\bar{x}) \leq f(x^*)$. Thus, $\bar{x} \in X^*$, and from the inequality $g(\bar{x}) \leq g(x^*)$ and the definition of x^* it then follows that $\bar{x} = x^*$. Thus $x_{\varepsilon_i} \to x^*$. But $\varepsilon \|x_\varepsilon - x^*\|^2 \leq \varepsilon(g(x^*) - g(x_\varepsilon))$, i.e., every sequence x_ε converges to x^*. □

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

$$g(x) = (B(x - a), x - a)/2 , \tag{12}$$

where $B > 0$. Then in the regularization method the quadratic function is minimized at each step, and therefore

$$x_\varepsilon = (A + \varepsilon B)^{-1}(b + \varepsilon B a) . \tag{13}$$

By (12) the matrix $A + \varepsilon B$ has an inverse for any $\varepsilon > 0$. Theorem 4 implies that $x_\varepsilon \to x^* \in X^*$, 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

$$x_\varepsilon = (A + \varepsilon I)^{-1} b . \tag{14}$$

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

$$f(x) = \|Cx - d\|^2 , \quad x \in \mathbf{R}^n , \tag{15}$$

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:

$$x^* = C^+ d , \tag{16}$$

where C^+ is some $n \times m$ matrix, referred to as the *pseudoinverse* of C. It follows from Theorem 4 and equality (14) that

$$C^+ = \lim_{\varepsilon \to +0} (C^T C + \varepsilon I)^{-1} C^T .$$

Other properties of pseudoinversion are given in Exercise 6 below.

Let us return to the regularization method. Clearly, the smaller ε, the closer x_ε to a solution, so that taking very small ε 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 ε. Here are examples to illustrate that $\|x_\varepsilon - x^*\|$ can be large even for small ε.

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 ε. 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

$$x^{k+1} = x_{\varepsilon_k} = \operatorname{argmin}\left(f(x) + \left(\frac{\varepsilon_k}{2}\right)\|x-a\|^2\right), \quad \varepsilon_k \to 0 .$$

We can try to go the different way: at each step, instead of the regularization parameter ε_k we vary the point a, but replace it with x^k. We thus arrive at the method

$$x^{k+1} = \operatorname*{argmin}_{x \in \mathbf{R}^n}\left(f(x) + \tfrac{1}{2}\varepsilon\|x - x^k\|^2\right), \quad \varepsilon > 0 , \tag{18}$$

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

$$\operatorname{Prox} a = \operatorname*{argmin}_{x \in \mathbf{R}^n}\left(f(x) + \tfrac{1}{2}\varepsilon\|x-a\|^2\right)$$

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:

$$x^{k+1} = \operatorname{Prox} x^k . \tag{19}$$

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

$$\psi(a) = \min_x\left[f(x) + \tfrac{1}{2}\varepsilon\|x-a\|^2\right]$$

is convex, differentiable, $\nabla\psi(a) = \varepsilon(a - \operatorname{Prox} a)$ satisfies a Lipschitz condition with constant ε and $X^* = \operatorname{Argmin}_a \psi(a) \neq \emptyset$. To minimize $\psi(a)$ we apply the gradient method with $\gamma = 1/\varepsilon$:

$$a^{k+1} = a^k - \varepsilon^{-1}\nabla\psi(a^k) = a^k - \varepsilon^{-1}\varepsilon(a^k - \operatorname{Prox} a^k) = \operatorname{Prox} a^k .$$

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 ε 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:

$$x^{k+1} = (A + \varepsilon I)^{-1}(b + \varepsilon x^k) . \tag{20}$$

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:

$$x^{k+1} = x^k - \gamma_k(\nabla f(x^k) + \varepsilon_k \nabla g(x^k)) , \tag{21}$$

where $g(x)$ is the regularization function, ε_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

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

$$\text{for all } x, \quad X^* = \operatorname*{Argmin}_{x \in \mathbf{R}^n} f(x) \neq \emptyset ,$$

and

$$0 \le \frac{\varepsilon_{k-1} - \varepsilon_k}{\varepsilon_k^2} \to 0, \quad 0 \le \varepsilon_k \to 0, \quad \sum_{k=0}^{\infty} \varepsilon_k = \infty, \tag{22}$$

$$\gamma_k = \gamma, \quad 0 < \gamma < \frac{2}{(1 + \varepsilon_0)L}. \tag{23}$$

Then in method (21), $x^k \to x^*$ where $x^* \in X^*$, $x^* = \operatorname{argmin}_{x \in X^*} g(x)$.

PROOF. Let $y^k \in \operatorname{argmin}_{x \in \mathbf{R}^n} \Phi_k(x)$, $\Phi_k(x) = f(x) + \varepsilon_k g(x)$; by the assumptions y^k exists, is uniquely defined and $y^k \to x^*$ (see Theorem 4). The function $\Phi_k(x)$ is strongly convex with constant $\ell\varepsilon_k$. Hence (see (35) of Section 1.1)

$$\Phi_k(y^{k-1}) \ge \Phi_k(y^k) + (\ell\varepsilon_k/2)\|y^k - y^{k-1}\|^2.$$

Similarly, from the strong convexity of $\Phi_{k-1}(x)$ we get

$$\Phi_{k-1}(y^k) \ge \Phi_{k-1}(y^{k-1}) + (\ell\varepsilon_{k-1}/2)\|y^k - y^{k-1}\|^2.$$

Adding these inequalities yields

$$(\varepsilon_k - \varepsilon_{k-1})(g(y^k) - g(y^{k-1})) + \ell(\varepsilon_{k-1} + \varepsilon_k)\|y^k - y^{k-1}\|^2/2 \le 0.$$

Since $\{y^k\}$ is bounded, there is a constant M such that

$$\|g(y^k) - g(y^{k-1})\| \le M\|y^k - y^{k-1}\|.$$

Hence

$$\|y^k - y^{k-1}\| \le \frac{2M(\varepsilon_{k-1} - \varepsilon_k)}{\ell(\varepsilon_{k-1} + \varepsilon_k)} \le N\frac{\varepsilon_{k-1} - \varepsilon_k}{\varepsilon_k}, \quad N = \frac{M}{\ell}. \tag{24}$$

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,

$$\ell\varepsilon_k I \le \nabla^2\Phi_k(x) \le L(1 + \varepsilon_k)I \le L(1 + \varepsilon_0)I.$$

Hence $\ell\varepsilon_k I \le A \le (1+\varepsilon_0)I$ and

$$\|x^{k+1} - y^k\| \le \|I - \gamma A\| \, \|x^k - y^k\| \tag{25}$$

$$\le \max_{\ell\varepsilon_k \le \lambda \le L(1+\varepsilon_0)} |1 - \gamma\lambda| \, \|x^k - y^k\| = (1 - \gamma\ell\varepsilon_k)\|x^k - y^k\|$$

for sufficiently large k as $\varepsilon_k \to 0$. Using (24) and (25), we get

$$\|x^{k+1} - y^k\| \le (1 - \gamma\ell\varepsilon_k)\|x^k - y^k\|$$

$$\le (1 - \gamma\ell\varepsilon_k)\|x^k - y^{k-1}\| + (1 - \gamma\ell\varepsilon_k)\|y^k - y^{k-1}\|$$

$$\le (1 - \gamma\ell\varepsilon_k)\|x^k - y^{k-1}\| + \mu_k ,$$

$$\mu_k = (1 - \gamma\ell\varepsilon_k)N(\varepsilon_{k-1} - \varepsilon_k)\varepsilon_k^{-1} .$$

Applying Lemma 3 of Section 2.2 for $u_k = \|x^k - y^{k-1}\|$ while taking (22) into account yields $u_k \to 0$. But $\|x^k - x^*\| \le \|x^k - y^{k-1}\| + \|y^{k-1} - x^*\| \to 0$ since $\|x^k - y^{k-1}\| \to 0$ by what was shown above, and $\|y^k - x^*\| \to 0$ by Theorem 4. □

With regard to the convergence rate, we see that by the condition $\sum_{k=0}^{\infty} \varepsilon_k = \infty$, the parameter ε_k cannot tend to zero too rapidly. On the other hand, the method converges not more rapidly that the method of regularization, and the latter, as we saw earlier, may converge slowly.

Exercises

5. Let $f(x)$ be a convex function in $\mathbf{R}^n$, $X^* = \operatorname{Argmin}_{x \in \mathbf{R}^n} f(x) \ne \emptyset$, let the function $g(x)$ be strictly convex, and let the set $\{x: g(x) \le \alpha\}$ be bounded and nonempty for some α. 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., $\|\text{Prox } a - \text{Prox } b\| \le \|a - b\|$;

c) the function

$$\psi(a) = \min_x (f(x) + (\tfrac{\varepsilon}{2})\|x-a\|^2)$$

is convex, differentiable, its gradient satisfies a Lipschitz condition with constant ε and is equal to $\nabla\psi(a) = \varepsilon(a - \text{Prox } a)$;

d) if

$$X^* = \operatorname{Argmin}_x f(x) \neq \emptyset ,$$

then

$$X^* = \operatorname{Argmin}_a \psi(a) .$$

8. Compute 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) 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) Prox $a = a$, $\psi(a) \equiv 0$;

c) Prox $a = [1 - 1/(\varepsilon\|a\|)]_+ a$, $\psi(a) = \varepsilon\|a\|^2/2$ for $\|a\| \leq 1/\varepsilon$,
$\psi(a) = \|a\| - 1/(2\varepsilon)$ for $\|a\| > 1/\varepsilon$.

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:

$$s^k = \nabla f(x^k) + r^k , \qquad \|r^k\| \leq \varepsilon . \tag{26}$$

In this situation, the gradient method (1) takes on the form

$$x^{k+1} = x^k - \gamma s^k . \tag{27}$$

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 ℓ and tends to infinity as $\ell \to 0$. 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 $\mathbf{R}^n$, the gradient of which satisfies a Lipschitz condition with constant L, and let $X^* = \operatorname{Argmin}_{x \in \mathbf{R}^n} f(x) \neq \emptyset$. Suppose the quantities L, ε 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

$$\|s^k\| \leq \varepsilon + 2\sqrt{\frac{\varepsilon L \rho}{2 - L\gamma}} \tag{28}$$

is satisfied, and let x_ε 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

$$\|x^{k+1} - x^*\| \leq \|x^k - x^*\| - \frac{\gamma(2 - L\gamma)\|\nabla f(x^k)\|^2}{2L\|x^k - x^*\|} + \gamma\varepsilon . \tag{29}$$

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

$$\|x^{k+1} - x^*\| \leq \|x^k - x^*\| - \gamma\varepsilon\left(\frac{2\rho}{\|x^k - x^*\|} - 1\right) . \qquad (30)$$

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 $\phi(\varepsilon) = O(\sqrt{\varepsilon}) \to 0$ as $\varepsilon \to 0$ and (2) this point is not too far away from the minimum point closest to the initial approximation. Using the inequality $f(x_\varepsilon) - f(x^*) \geq \|\nabla f(x_\varepsilon)\| \, \|x_\varepsilon - x^*\|$ one can guarantee that at x_ε the value of the function is also close to the minimal value:

$$f(x_\varepsilon) - f(x^*) = O(\sqrt{\varepsilon}) . \qquad (31)$$

In this sense, the point x_ε 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_ε 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 γ_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 $\Sigma_{k=0}^{\infty} \varepsilon_k < \infty$ 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 δ. Let

$$\Phi_\varepsilon(x_\varepsilon^\delta) \leq \Phi_\varepsilon^* + \delta, \tag{32}$$

where

$$\Phi_\varepsilon(x) = f(x) + \varepsilon g(x), \quad \Phi_\varepsilon^*(x) = \min_{x \in \mathbf{R}^n} \Phi_\varepsilon(x).$$

THEOREM 8. Let the conditions of Theorem 4 be satisfied. Then as

$$\varepsilon \to 0, \quad \delta/\varepsilon \to 0 \tag{33}$$

one has $x_\varepsilon^\delta \to x^*$.

PROOF. Is the same as that of Theorem 4. □

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 $\mathbf{R}^n$, where $\nabla f(x)$ satisfies a Lipschitz condition, $X^* = \operatorname{Argmin}_{x \in \mathbf{R}^n} f(x) \neq \emptyset$. Let $s^k = \nabla f(x^k) + \xi^k$, where the random noise ξ^k is independent and $\mathbf{E}\xi^k = 0$, $\mathbf{E}\|\xi^k\|^2 \leq \sigma^2$. Consider the gradient method $x^{k+1} = x^k - \gamma_k s^k$ under the conditions $\Sigma_{k=0}^\infty \gamma_k = \infty$, $\Sigma_{k=0}^\infty \gamma_k^2 < \infty$. Using the method of proof of Theorem 1 of this section and Theorem 1 of Section 2.2, prove that $x^k \to \bar{x} \in X^*$ 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, $\Sigma_{i=1}^{n+1} x_i t^{i-1}$. After the appropriate norm (L_1, L_2, L_∞ and such) is chosen, the problem reduces to minimizing the function $f(x) = \|a(t) - \Sigma_{i=1}^{n+1} x_i t^{i-1}\|$. 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_ε such that $\|x_\varepsilon - x^*\| \leq \|a - x^*\|$, 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

$$\min f(x), \quad x \in \mathbf{R}^n, \tag{1}$$

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 $\nabla f(x^*) = 0$ 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 $\lambda_1 \leq \cdots \leq \lambda_n$ being its eigenvalues and $e^1, \ldots, e^n$ 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 $y \in L_- = \{\Sigma_{i:\lambda_i<0} \gamma_i e^i\}$, $y \neq 0$, are directions of decrease and the vectors $y \in L_+ = \{\Sigma_{i:\lambda_i>0} \gamma_i e^i\}$, $y \neq 0$, 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_+. □

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

$$\begin{aligned} x^{k+1} - x^* &= x^k - x^* - \gamma \nabla f(x^k) \\ &= (I - \gamma \nabla^2 f(x^*))(x^k - x^*) + o(x^k - x^*) . \end{aligned} \tag{2}$$

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 $1 - \gamma\lambda_i$, $i = 1, \ldots, n$, but all $\lambda_i \neq 0$, 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 leaves this neighborhood (except for the special case where $x^0 = x^*$).

For the case where x^* is a saddle point, the analysis is simple if $f(x)$ is a quadratic function. Then

$$x^{k+1} - x^* = (I - \gamma A)(x^k - x^*) , \quad x^k - x^* = (I - \gamma A)^k (x^0 - x^*) , \tag{3}$$

$A = \nabla^2 f(x)$. In the notation of Theorem 1, $(x^k - x^*, e^i) = (1 - \gamma\lambda_i)^k \times (x^0 - x^*, e^i)$. If $\lambda_i > 0$, $0 < \gamma < 2/\|A\|$, then $|1 - \gamma\lambda_i| < 1$, and so $(x^k - x^*, e^i) \to 0$. 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 $(x^k - x^*, e^i) \to \infty$ for $(x^0 - x^*, e^i) \neq 0$. Since $\|x^k - x^*\|^2 = \sum_{i=1}^n (x^k - x^*, e^i)^2$, we get $\|x^k - x^*\| \to \infty$ 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.

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 $Q = \{a \le x \le b\} \subset \mathbf{R}^n$, and let x^k be the sequence of independent uniformly distributed random vectors on Q. Then

$$\min_{1 \le i \le k} f(x^i) \xrightarrow{P} \min_{x \in Q} f(x) .$$

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 $f(x) \le f(x^*) + \varepsilon$ 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^i being in $U \cap Q$ is v/V; the probability that at least one of the points $x^1, \ldots, x^k$ is in $U \cap Q$ is $p_k = 1 - (1 - v/V)^k$. Obviously, $p_k \to 1$ as $k \to \infty$, i.e.,

$$\mathbf{P}\left\{ \min_{1 \le i \le k} f(x^i) > f(x^*) + \varepsilon \right\} \to 0 \quad \text{as} \quad k \to \infty .$$

This implies the convergence in probability. □

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 $x = (x_1, \ldots, x_{10}) \in \mathbf{R}^{10}$, $f(x) = \max_{1 \le j \le 10} x_j$, $Q = \{x: 0 \le x_j \le 1, j = 1, \ldots, 10\}$, 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

$$|f(x) - f(y)| \leq L\|x - y\| , \tag{4}$$

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: $\phi_{k-1} = \min\limits_{1 \leq i \leq k-1} f(x^i)$ 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 Lipshitz condition:

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \tag{5}$$

and L is known, the region can be diminished by means of the inequality

$$|f(x) - f(x^k) - (\nabla f(x^k), x - x^k)| \leq (L/2)\|x - x^k\|^2 . \tag{6}$$

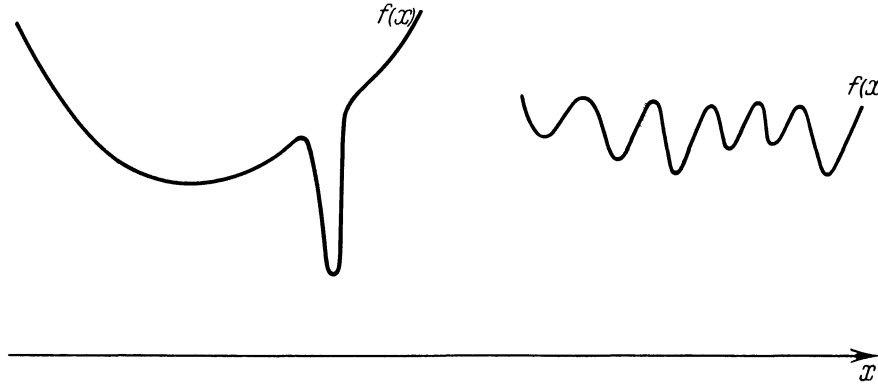


Fig. 26 Functions for which the global minimum is difficult to find.

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

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}) . \qquad (7)$$

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 μ (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

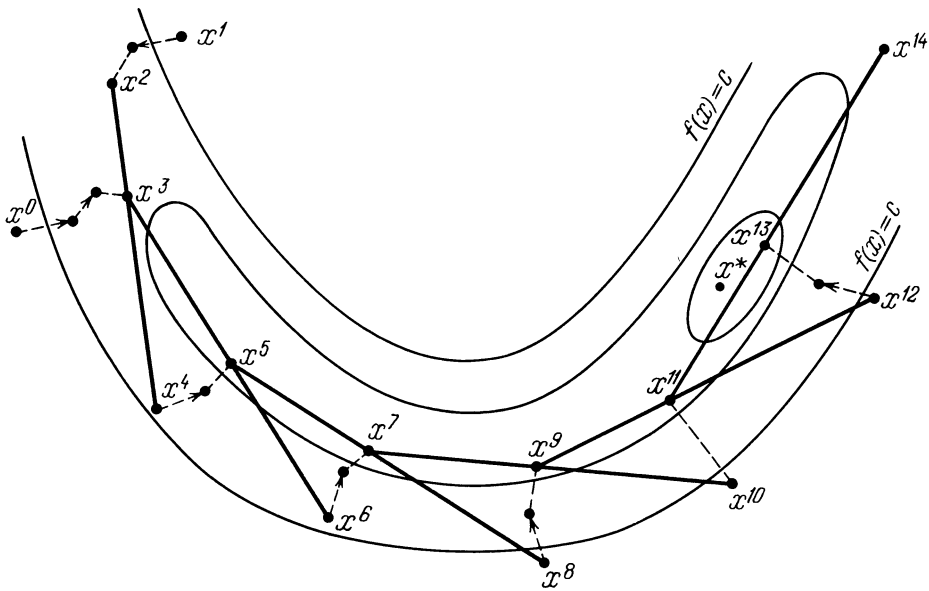


Fig. 27 The gully method.

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) 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.

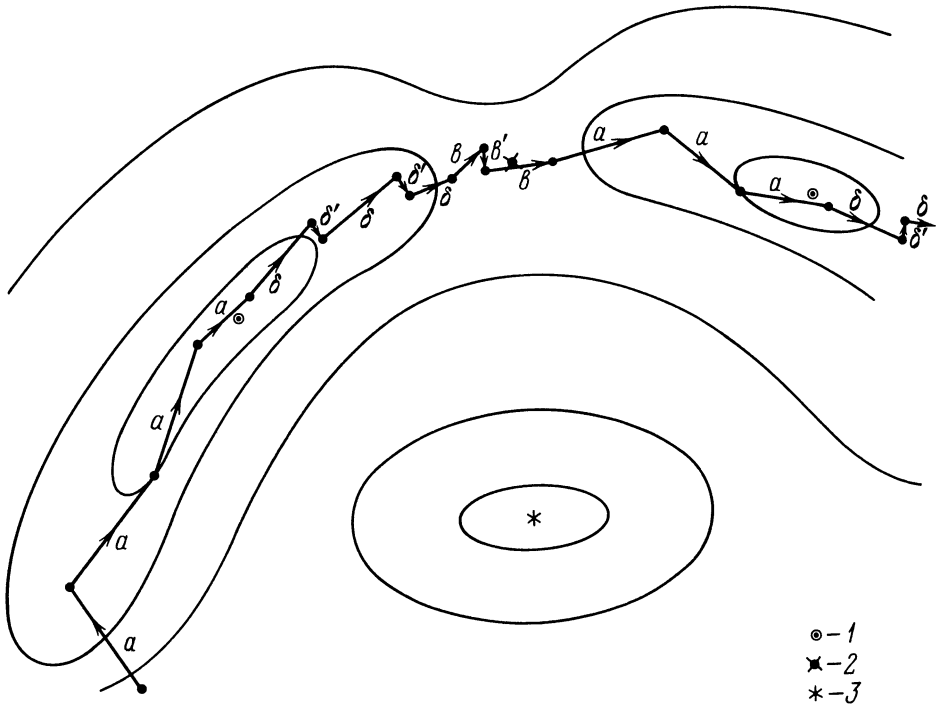


Fig. 28 The DAS method for global minimization: a is the descent stage; δ is the ascent stage; β is the saddle point; 1 denotes minimum points; 2 denotes the saddle point; 3 denotes the maximum point.

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 $Q = \{x: 0 \le x_i \le 1\}$. 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 "naîve" 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, ..., 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 $f(x^1)$, ..., $f(x^k)$. 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 $f(x, t)$, $x \in \mathbf{R}^n$, $t \in \mathbf{R}^1$. We write the local or global minimum of $f(x, t)$ for fixed $t = t_0$ as

$$x_0^* = \underset{x \in \mathbf{R}^n}{\operatorname{argmin}} f(x, t_0) . \tag{1}$$

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

$$\frac{dx^*}{dt} = -[\nabla_{xx}^2 f(x^*(t), t)]^{-1} \nabla_{xt}^2 f(x^*(t), t) , \qquad x^*(t_0) = x_0^* . \tag{2}$$

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.,

$$x^*(t) = \underset{x \in \mathbf{R}^n}{\operatorname{argmin}} f(x, t) . \tag{3}$$

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 $\nabla^2_{xx} f(x, t)$ and $\nabla^2_{tx} f(x, t)$ 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 kth 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

$$x^{k+1} = x^k - \gamma \nabla f_k(x^k), \tag{4}$$

or, of Newton's method

$$x^{k+1} = x^k - [\nabla^2 f_k(x^k)]^{-1} \nabla f_k(x^k). \tag{5}$$

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 nonstationary 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 $f_k(x) \to f(x)$, $\nabla f_k(x) \to \nabla f(x)$, 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

$$lI \leq \nabla^2 f_k(x) \leq LI, \quad l > 0, \tag{6}$$

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:

$$\|x_k^* - x_{k+1}^*\| \leq a . \tag{7}$$

THEOREM 1. Under the assumptions made above, for method (4) with $0 < \gamma < 2/L$ we have the bound

$$\overline{\lim_{k\to\infty}} \|x^k - x_k^*\| \leq \frac{a}{1-q}, \quad q = \max\{|1-\gamma\ell|, |1-\gamma L|\} < 1 . \tag{8}$$

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). □

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

$$x_{k+1}^* = g_k(x_k^*) , \tag{9}$$

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

$$x^{k+1} = g_k(x^k) - \gamma\nabla f_k(g_k(x^k)) . \tag{10}$$

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

$$\min_{x \in Q \subset \mathbf{R}^n} f(x), \tag{A}$$

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 $Q = \{x: a \le x \le b\}$, the ball $Q = \{x: \|x\| \le \alpha\}$, 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 $f(x) \ge f(x^*)$ for all $x \in Q$, $\|x - x^*\| \le \varepsilon$ for some $\varepsilon > 0$. If $f(x) \ge f(x^*)$ 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

$$(\nabla f(x^*), x - x^*) \ge 0 \quad \text{for all } x \in Q. \tag{1}$$

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 $0 \le \alpha \le 1$ 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^*. □

A vector $a \in \mathbf{R}^n$ satisfying the condition $(a, x - x^*) \le 0$ for all $x \in Q$ is said to support Q at the point $x^* \in Q$ (if $a \ne 0$, 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 $0 \le \alpha \le 1$. 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

$$(\nabla f(x^*), x - x^*) \ge \alpha \|x - x^*\|, \quad \alpha > 0, \tag{2}$$

be satisfied for all $x \in Q$, $\|x - x^*\| \le \varepsilon$, $\varepsilon > 0$. Then x^* is a local minimum point of $f(x)$ on Q.

PROOF. Take $\varepsilon_1 > 0$, $\varepsilon_1 \le \varepsilon$, so that

$$|f(x^* + y) - f(x^*) - (\nabla f(x^*), y)| \le \alpha \|y\|/2$$

for $\|y\| \le \varepsilon_1$. Then for $x \in Q$, $\|x - x^*\| \le \varepsilon_1$, we have

$$f(x) \ge f(x^*) + (\nabla f(x^*), x - x^*) - \alpha \|x - x^*\|/2 \ge f(x^*) + \alpha \|x - x^*\|/2,$$

i.e., x^* is a local minimum point. □

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

$$f'(x^*; s) \geq \alpha \|s\| , \qquad \alpha > 0 , \tag{3}$$

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

$$Q = \{x \in \mathbf{R}^n : a \leq x \leq b\} . \tag{4}$$

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

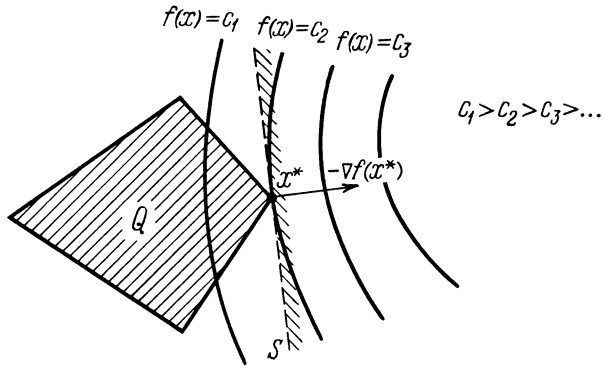


Fig. 29 Extremum conditions on the set Q.

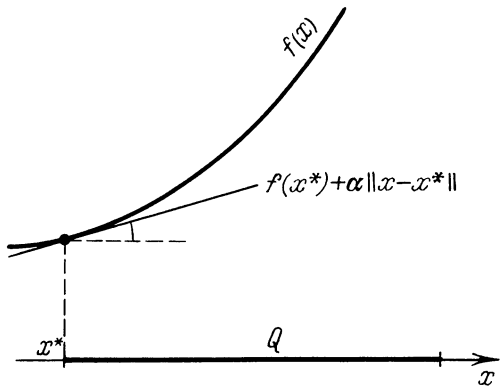


Fig. 30 The sharp minimum in a constrained problem.

The page is omited (the following page was omited in hardcopy)

$$\frac{\partial f(x^*)}{\partial x_i}\begin{cases} = 0, & a_i < x_i^* < b_i\,, \\ \geq 0, & x_i^* = a_i\,, \\ \leq 0, & x_i^* = b_i\,, \end{cases} \tag{5}$$

and if $x_i^* = a_i$ or $x_i^* = b_i$ for all $1 \leq i \leq n$ and

$$\frac{\partial f(x^*)}{\partial x_i}\begin{cases} > 0, & x_i^* = a_i\,, \\ < 0, & x_i^* = b_i\,, \end{cases} \tag{6}$$

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 $I = \{1, \ldots, n\}$ 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 relay on extremum conditions.

As second example let's consider a minimization problem on a linear manifold

$$Q = \{x \in \mathbf{R}^n : Ax = b\}\,, \tag{7}$$

where $b \in \mathbf{R}^m$, A is a matrix $m \times n$. From Theorem 1 it follows that $\left(\nabla f(x^*),\, x - x^*\right) \geq 0$ for all $x \in Q$, i.e. $\left(\nabla f(x^*),\, s\right) \geq 0$ for all $s \in L = \{s : As = 0\}$. If there exists $s^0 \in L$ such that $\left(\nabla f(x^*),\, s^0\right) > 0$, then $\left(\nabla f(x^*),\, -s^0\right) < 0$, which is a contradiction with $-s^0 \in L$. Thus $\left(\nabla f(x^*),\, s\right) = 0$ for all $s \in L$. From here follows that (Lemma 1, Section 1, Chapter 8) there exists $y^* \in \mathbf{R}^m$ such that

$$\nabla f(x^*) = A^T y^*\,. \tag{8}$$

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 $x \in Q$,

$$(\partial f(x^*), x - x^*) \geq 0 , \tag{9}$$

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 Γ be the closure of the cone generated by the feasible directions, i.e., $\Gamma = \{x: x = \lim_{k\to\infty} \lambda_k(x^k - x^*), \lambda_k > 0, x^k \in Q\}$. *If* $c \notin \Gamma$, then we apply again the separation theorem (this is possible since Γ 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 Γ 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 $\lambda_k(x^k - x^*) \to c$. Take k such that

$$\|\lambda_k(x^k - x^*) - c\| \leq \alpha/(2L) , \qquad L = \max_{a \in S} \|a\| .$$

Then by Lemma 6 of Section 5.1,

$$f'(x^*; \lambda_k(x^k - x^*)) = \max_{a \in S} (a, \lambda_k(x^k - x^*))$$

$$= \max_{a \in S} (a, c) + \max_{a \in S} (\lambda_k(x^k - x^*) - c, a) \leq -\alpha + \tfrac{1}{2}\alpha = -\tfrac{1}{2}\alpha .$$

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. □

Exercise

2. Using Theorem 3, show that $b = P_Q(a)$ iff $(b-a, x-b) \geq 0$ for all $x \in Q$ (cf. (5) Section 5.1). *Hint*: $P_Q(a)$ is a solution of the problem min $\|x-a\|^2$, $x \in Q$.

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 nonempty for some α. Then problem (A) has a solution. □

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

$$f(x) \geq f(x^*) + \alpha\|x - x^*\|/2, \qquad \|x - x^*\| \leq \varepsilon, \tag{10}$$

derived in proving Theorem 2. □

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. □

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 $x^k \in Q$, $f(x^k) \to f^* = \inf_{x \in Q} f(x)$ that $x^k \to x^*$, $f(x^*) = f^*$. One can define the generalized minimizing sequence: $f(x^k) \to f^* = \inf_{x \in Q} f(x)$, $\rho(x^k, Q) \to 0$, 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. □

One can obtain quantative 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

$$f(x) \geq f(x^*) + \alpha \|x - x^*\|, \quad \alpha > 0 \tag{11}$$

(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. □

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^*. □

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 α.

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 ℓ 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)

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

where P_Q is the projector onto Q (see Section 5.1).

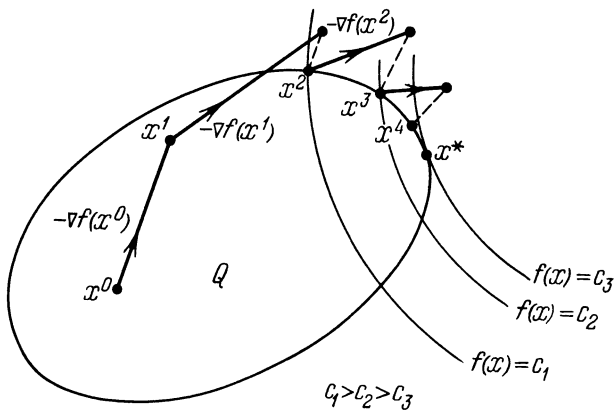


Fig. 31 The gradient-projection method.

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, $x^* = \operatorname{Argmin}_{x \in Q} f(x) \neq \emptyset$ and $0 < \gamma < 2/L$. Then

(i) $x^k \to x^* \in X^*$;

(ii) if $f(x)$ is strongly convex, then $x^k \to x^*$ with the rate of geometric progression;

(iii) if $f(x)$ is twice differentiable and $\ell I \leq \nabla^2 f(x) \leq LI$, $x \in Q$, $\ell > 0$, 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 $\bar{x}$ be an arbitrary minimum point. Then taking $x = x^k - \gamma\nabla f(x^k)$, $y = \bar{x}$ in (5) of Section 5.1, one obtains

$$(x^k - \gamma\nabla f(x^k) - x^{k+1}, \bar{x} - x^{k+1}) \leq 0 .$$

Using the minimum condition (1) of Section 7.1 yields

$$0 \geq (x^k - x^{k+1}, \bar{x} - x^{k+1}) - \gamma(\nabla f(x^k) - \nabla f(\bar{x}), \bar{x} - x^{k+1}) .$$

Let us transform the right-hand terms:

$$(x^k - x^{k+1}, \bar{x} - x^{k+1}) = (\|x^k - \bar{x}\|^2 - \|x^k - x^{k+1}\|^2 - \|x^{k+1} - \bar{x}\|^2)/2 ,$$

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

Here inequality (11) of Section 1.1 was applied first and the inequality $\|a\|^2 + (a,b) \geq -\|b\|^2/4$ next, using $a = L^{-1/2}(\nabla f(x^k) - \nabla f(\bar{x}))$, $b = L^{1/2} \times (x^{k+1} - x^k)$. Hence

$$0 \geq \|x^k - \bar{x}\|^2/2 - (1 - L\gamma/2)\|x^{k+1} - x^k\|^2/2 - \|x^{k+1} - \bar{x}\|^2/2 ,$$

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

This implies that the limit $\lim_{k\to\infty} \|x^k - \bar{x}\|$ exists and $\lim_{k\to\infty} \|x^{k+1} - x^k\| = 0$. 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 $\bar{x} = x^*$, 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

$$\begin{aligned}\|T(x) - T(y)\|^2 &\le \|x - \gamma\nabla f(x) - y + \gamma\nabla f(y)\|^2 \\ &= \|x - y\|^2 - 2\gamma(\nabla f(x) - \nabla f(y), x - y) + \gamma^2\|\nabla f(x) - \nabla f(y)\|^2 \\ &\le \|x - y\|^2 - \gamma(2 - \gamma L)(\nabla f(x) - \nabla f(y), x - y) \le q^2\|x - y\|^2 ,\end{aligned}$$

$$q^2 = 1 - \gamma\ell(2 - \gamma L) < 1 .$$

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

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

$$A_k = \int_0^1 \nabla^2 f(x^* + \tau(x^k - x^*))\, d\tau , \qquad q = \max\{|1 - \gamma\ell|, |1 - \gamma L|\}$$

(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

$$\begin{aligned}&(x^k - \gamma\nabla f(x^k) - x^*, x - x^*) \\ &\quad = (x^k - x^* - \gamma(\nabla f(x^k) - \nabla f(x^*)), x - x^*) - \gamma(\nabla f(x^*), x - x^*) \\ &\quad \le (1 + \gamma L)\|x^k - x^*\|\,\|x - x^*\| - \gamma\alpha\|x - x^*\| \\ &\quad = ((1 + \gamma L)\|x^k - x^*\| - \gamma\alpha)\|x - x^*\| \le 0\end{aligned}$$

for $\|x^k - x^*\| \le \gamma\alpha/(1 + \gamma L)$. Since $x^k \to x^*$, 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^*. □

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

$$x^{k+1} = (x^k - \gamma \nabla f(x^k))_+ . \tag{2}$$

Let $Q = \{x: a \leq x \leq b\}$, $x \in \mathbf{R}^n$. For scalars τ, $\alpha \leq \beta$, we set

$$(\tau)_\alpha^\beta = \begin{cases} \tau , & \alpha \leq \tau \leq \beta , \\ \beta , & \tau > \beta , \\ \alpha , & \tau < \alpha . \end{cases} \tag{3}$$

The notation $(x)_a^b$ has an analogous meaning for the vector x, $a \leq b$; this is the vector whose ith component is equal to $(x_i)_{a_i}^{b_i}$. Then the gradient projection method for the given Q has the form:

$$x^{k+1} = (x^k - \gamma \nabla f(x^k))_a^b . \tag{4}$$

Further, let Q be the ball $Q = \{x: \|x\| \leq \rho\}$. Then

$$x^{k+1} = \begin{cases} x^k - \gamma \nabla f(x^k) & \text{if } \|x^k - \gamma \nabla f(x^k)\| \leq \rho , \\ \rho \dfrac{x^k - \gamma \nabla f(x^k)}{\|x^k - \gamma \nabla f(x^k)\|} & \text{if } \|x^k - \gamma \nabla f(x^k)\| > \rho . \end{cases} \tag{5}$$

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

$$x^{k+1} = (I - C^+ C)(x^k - \gamma \nabla f(x^k)) + C^+ d . \tag{6}$$

Here C^+ is the pseudoinverse of C (Section 6.1). If $x^0 \in Q$, then

$$x^{k+1} - x^0 = T(x^k - x^0 - \gamma \nabla f(x^k)) , \tag{7}$$

where $T = I - C^+ C$, whereas if C is a matrix of rank $m < n$, then $T = I - C^T (CC^T)^{-1} C$.

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 $\|x^{k+1} - x^k\| \to 0$.
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

$$x^{k+1} = P_Q(x^k - \gamma_k \partial f(x^k)), \tag{8}$$

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 γ_k are analogous to those considered in Section 5.3, and we shall mention only two very important ones:

$$\gamma_k \to 0, \quad \sum_{k=0}^{\infty} \gamma_k = \infty, \tag{9}$$

$$\gamma_k = \frac{f(x^k) - f^*}{\|\partial f(x^k)\|^2}, \quad f^* = \min_{x \in Q} f(x). \tag{10}$$

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. □

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^k 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:

$$\begin{aligned} \bar{x}_k &= \operatorname*{argmin}_{x \in Q} (\nabla f(x^k), x), \\ x^{k+1} &= x^k + \gamma_k(\bar{x}^k - x^k). \end{aligned} \tag{11}$$

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 γ_k, $0 \le \gamma_k \le 1$.

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 γ_k be defined from the steepest descent condition:

$$\gamma_k = \operatorname*{argmin}_{0 \le \gamma \le 1} f(x^k + \gamma(\bar{x}^k - x^k)). \tag{12}$$

Then

(i) $(\nabla f(x^k), x^k - \bar{x}^k) \to 0$ 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:

$$f(x^k) - f^* = O(1/k), \quad f^* = \min_{x \in Q} f(x), \tag{13}$$
$$f(x^k) \ge f^* \ge f(x^k) + (\nabla f(x^k), \bar{x}^k - x^k);$$

(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

$$V(x^{k+1}) = \min_{0 \le \gamma \le 1} V(x^k - \gamma s^k) \le \min_{0 \le \gamma \le 1} \phi(\gamma),$$

$$\phi(\gamma) = V(x^k) - \gamma(\nabla f(x^k), s^k) + \gamma^2 L \|s^k\|^2/2.$$

Set $\gamma_k^* = (\nabla f(x^k), s^k)/L\,\|s^k\|^2$. By the definition of $\bar{x}^k$: $(\nabla f(x^k), s^k) \ge 0$, i.e., $\gamma_k^* \ge 0$. Two cases are possible: 1) $\gamma_k^* \le 1$ and 2) $\gamma_k^* > 1$. In case 1:

$$V(x^{k+1}) \le \phi(\gamma_k^*) \le V(x^k) - \frac{(\nabla f(x^k), s^k)^2}{2L\|s^k\|^2} \le V(x^k) - \frac{(\nabla f(x^k), s^k)^2}{2LR^2}, \tag{14}$$

where R is the diameter of the set Q. In case 2: $L\|s^k\|^2 < (\nabla f(x^k), s^k)$ and

$$V(x^{k+1}) \le \phi(1) \le V(x^k) - (\nabla f(x^k), s^k) + (L/2)\|s^k\|^2 \le V(x^k) - (\nabla f(x^k), s^k)/2. \tag{15}$$

Thus in both cases $V(x^k)$ is monotonically decreasing and since $V(x) \geq 0$, then $V(x^k) - V(x^{k+1}) \to 0$. By (14) and (15) this implies that $(\nabla f(x^k), s^k) \to 0$.

Now let x^* be any limit point of the sequence x^k (it is known to exist since Q is bounded), $x^{k_i} \to x^*$. Then for any $x \in Q$,

$$(\nabla f(x^*), x-x^*) = (\nabla f(x^*) - \nabla f(x^{k_i}), x-x^*) + (\nabla f(x^{k_i}), x-\bar{x}^{k_i}) + (\nabla f(x^{k_i}), \bar{x}^{k_i}-x^{k_i}) + (\nabla f(x^{k_i}), x^{k_i}-x^*) .$$

The first and fourth terms on the right tend to zero as $i \to \infty$ since $x^{k_i} \to x^*$, 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, $V(x^k) \to 0$ 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

$$(\nabla f(x^k), s^k) \leq \max \{\left[2LR^2(V(x^k) - V(x^{k+1}))\right]^{1/2}, 2(V(x^k) - V(x^{k+1}))\}$$

$$\leq \left[2LR^2(V(x^k) - V(x^{k+1}))\right]^{1/2}$$

for sufficiently large k since $V(x^k) \to 0$. 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

$$(\nabla f(x^k), x^*-x) = (\nabla f(x^*), x^*-x) + (\nabla f(x^k) - \nabla f(x^*), x^*-x)$$

$$\leq -\alpha\|x-x^*\| + L\|x^k-x^*\| \, \|x-x^*\| \leq 0$$

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 $x^{k+1} = x^*$. □

Let us illustrate with an example that the bound (13) cannot be improved even for a strongly convex $f(x)$. Let $x \in \mathbf{R}^2$,

$$f(x) = x_1^2 + (1 + x_2)^2 , \quad Q = \{x: |x_1| \leq 1, 0 \leq x_2 \leq 1\} \qquad (16)$$

(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 $v_{k+1} = v_k - (\frac{1}{4})\|s^k\|^2 v_k^2$, $\|s^k\|^2 \to 1$, 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 γ_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 nonsmooth 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.

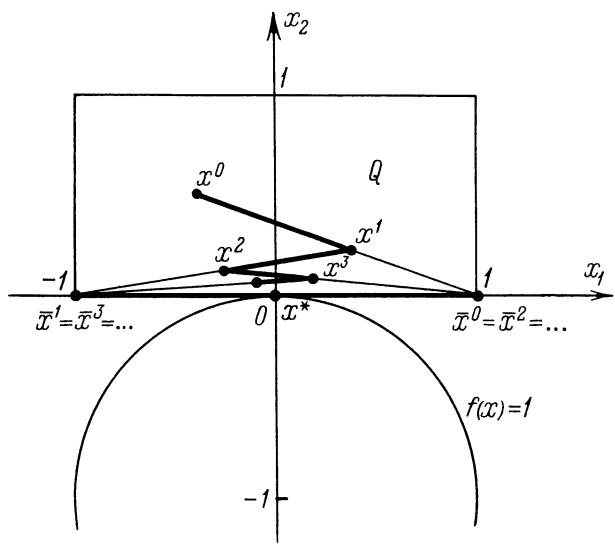


Fig. 32 Slow convergence of the ~~constrained~~ conditional gradient method.

7. Verify that the solution of the following elementary minimization problem is correct:

$$\text{(a)}\quad x^* = \operatorname*{argmin}_{a \le x \le b} (c, x),\quad x_i^* = \begin{cases} a_i & \text{if } c_i > 0\ , \\ b_i & \text{if } c_i < 0\ , \\ \text{anything between } a_i \text{ and } b_i & \text{if } c_i = 0\ ; \end{cases}$$

$$\text{(b)}\quad x^* = \operatorname*{argmin}_{\|x\| \le \rho} (c, x) = -\frac{c\rho}{\|c\|}\ ;$$

$$\text{(c)}\quad x^* = \operatorname*{argmin}_{(1/2)(Ax,x)-(b,x) \le \alpha} (c, x),\ A > 0,\ \alpha > 0,\ x^* = A^{-1}(b - \lambda c),$$

where λ 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 $\lambda \to \infty$.

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

$$\|z - (x+y)/2\| \le \beta\|x - y\|^2\ .$$

Show that if $f(x)$ is a strongly convex function on $\mathbf{R}^n$, then the set $Q_\alpha = \{x: f(x) \le \alpha\}$ 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 $\|\nabla f(x)\| \ge \varepsilon > 0$ 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

$$\psi(x) = \min_{y \in Q} [f(x) + (\nabla f(x), y - x)]\ .$$

Show that if $f(x)$ is convex, then $\psi(x) \le f(x)$ for all $x \in Q$, and equality obtains iff $x = \operatorname{argmin}_{x' \in Q} f(x')$. 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-

mum. 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

$$x^{k+1} = \arg\min_{x \in Q} f_k(x) ,$$
$$f_k(x) = f(x^k) + (\nabla f(x^k), x-x^k) + (\nabla^2 f(x^k)(x-x^k), x-x^k)/2 . \tag{17}$$

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

$$\nabla^2 f(x^*) > 0 . \tag{18}$$

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

$$0 \leq (\nabla f(x^k) + \nabla^2 f(x^k)(x^{k+1}-x^k), x^*-x^{k+1})$$
$$= (\nabla f(x^*), x^*-x^{k+1})$$
$$+ (\nabla f(x^k) - \nabla f(x^*) + \nabla^2 f(x^k)(x^{k+1}-x^k), x^*-x^{k+1}) .$$

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

$$\nabla f(x^k) - \nabla f(x^*) = \nabla^2 f(x^k)(x^k-x^*) + r , \qquad \|r\| \leq (L/2)\|x^k-x^*\|^2$$

(see (15) of Section 1.1). Hence

$$0 \leq (\nabla^2 f(x^k)(x^k-x^*) + r + \nabla^2 f(x^k)(x^{k+1}-x^k), x^*-x^{k+1})$$
$$\leq -\ell \|x^{k+1}-x^*\|^2 + (L/2)\|x^k-x^*\|^2 \|x^{k+1}-x^*\| .$$

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

$$\|x^{k+1} - x^*\| \leq L\|x^k - x^*\|^2/(2\ell) . \tag{19}$$

If $L\|x_0 - x^*\|/(2\ell) < 1$, 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. □

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 λ 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

$$x^{k+1} = \operatorname*{argmin}_{x \in Q} ((\nabla f(x^k), x - x^k) + \tfrac{1}{2} (H_k(x - x^k), x - x^k)) , \tag{1}$$

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

$$x^{k+1} = \operatorname*{argmin}_{x \in Q} \|x - (x^k - \gamma \nabla f(x^k))\|^2 .$$

It is possible to extend the class of methods (1), introducing the one-dimensional procedure:

$$\begin{aligned} \bar{x}^k &= \operatorname*{argmin}_{x \in Q} ((\nabla f(x^k), x - x^k) + \tfrac{1}{2}(H_k(x - x^k), x - x^k)) , \\ x^{k+1} &= x^k + \gamma_k s^k , \quad s^k = \bar{x}^k - x^k , \quad \gamma_k = \operatorname*{argmin}_{0 \le \gamma \le 1} f(x^k + \gamma s^k) . \end{aligned} \tag{2}$$

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 $\ell I \le \nabla^2 f(x) \le LI$, $\ell > 0$ for all $x \in Q$, Q being closed and convex, and

$$\|H_k - \nabla^2 f(x^k)\| \le \varepsilon < \ell/2 . \tag{3}$$

Then in method (1) the x^k converges locally to $x^* = \operatorname*{argmin}_{x \in Q} f(x)$ with the rate of geometric progression, whereas if

$$\|H_k - \nabla^2 f(x^k)\| \to 0 , \tag{4}$$

the rate is superlinear.

PROOF. From the definition of x^{k+1} we have

$$\begin{aligned} 0 &\le (\nabla f(x^k) + H_k(x^{k+1} - x^k), x^* - x^{k+1}) \\ &\le (\nabla f(x^k) - \nabla f(x^*) + H_k(x^{k+1} - x^k), x^* - x^{k+1}) . \end{aligned}$$

But

$$\nabla f(x^k) - \nabla f(x^*) = \nabla^2 f(x^k)(x^k - x^*) + r, \qquad \|r\| \le \tfrac{1}{2} L \|x^k - x^*\|^2,$$

and hence

$$\begin{aligned} 0 &\le ((\nabla^2 f(x^k) - H_k)(x^k - x^*) + H_k(x^{k+1} - x^*) + r, x^* - x^{k+1}) \\ &\le \varepsilon \|x^k - x^*\| \, \|x^{k+1} - x^*\| - (\ell - \varepsilon)\|x^{k+1} - x^*\|^2 \\ &\qquad + (L/2)\|x^k - x^*\|^2 \|x^{k+1} - x^*\|, \end{aligned}$$

$$\|x^{k+1} - x^*\| \le \frac{\varepsilon + (L/2)\|x^k - x^*\|}{\ell - \varepsilon} \|x^k - x^*\|.$$

If $(L/2)\|x^0 - x^*\| < \ell - 2\varepsilon$, then $x^k \to x^*$ with the rate of geometric progression: the smaller ε the smaller the ratio. Similarly, by (4) we have

$$\|x^{k+1} - x^*\| \le q_k \|x^k - x^*\|, \qquad q_k \to 0 \quad \text{as } k \to \infty,$$

which implies superlinear convergence. □

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

$$\nabla f(x^{i+1}) - \nabla f(x^i) \approx H(x^{i+1} - x^i), \qquad i = k, \ldots, k-n+1, \tag{5}$$

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

$$x^{k+1} = P_Q(x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k))$$

does not generally converge. In particular, if $f(x)$ is quadratic, then for any x^0 the method stops at x^1 (identical for all x^0), and x^1 is not generally a solution.

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 $\mathbf{R}^n$, $Q = \{x: Cx = 0\}$, C is an $m \times n$ matrix of rank m. The projection of a vector onto this subspace is given by

$$P_Q(x) = (I - C^+ C)x = (I - C^T(CC^T)^{-1}C)x .$$

Let us write the conjugate gradient method in which the vector $\nabla f(x)$ is replaced by its projection onto Q:

$$\begin{aligned} x^{k+1} &= x^k + \alpha_k p^k , \quad \alpha_k = \operatorname{argmin}_{\alpha} f(x^k + \alpha p^k) , \quad x^0 \in Q , \\ p^k &= -P_Q \nabla f(x^k) + \beta_k p^{k-1} , \quad p^0 = -P_Q \nabla f(x^0) , \qquad (6) \\ \beta_k &= \|P_Q \nabla f(x^k)\|^2 / \|P_Q \nabla f(x^{k-1})\|^2 . \end{aligned}$$

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 $\mathbf{R}^n$, 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 $\{1,...,n\}$. 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:

$$x^{k+1} = x^k + \alpha_k p^k, \quad \alpha_k = \underset{\substack{\alpha \geq 0 \\ x^k + \alpha p^k \geq 0}}{\operatorname{argmin}} f(x^k + \alpha p^k),$$

$$p_i^k = \begin{cases} -\nabla f(x^k)_i + \beta_k p_i^{k-1}, & i \in I_k, \\ 0, & i \in I_k, \end{cases}$$

$$\beta_k = \begin{cases} \Sigma_{i \in I_k} (\nabla f(x^k)_i)^2 / \Sigma_{i \in I_k} (\nabla f(x^{k-1})_i)^2 & \text{if } I_k = I_{k-1}, \\ 0 & \text{if } k = 0 \text{ or } I_k \neq I_{k-1}, \end{cases}$$

$$I_k = \begin{cases} \{i: x_i^k = 0, \nabla f(x^k)_i > 0\} & \text{if } k = 0 \text{ or } \nabla f(x^k)_i = 0 \\ & \text{for all } i \in I_{k-1}, \\ I_{k-1} \cup \{i: i = 0\} & \text{otherwise}. \end{cases}$$

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

$$\| \tilde{\nabla} f(x^k) - \nabla f(x^k) \| \leq \varepsilon \qquad (\| \partial f(x^k) - \tilde{\partial} f(x^k) \| \leq \varepsilon) . \tag{1}$$

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 ε. The situation is slightly different with the conditional gradient method. First of all it includes the one-dimensinal 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. □

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

$$\xi^k = \nabla f(x^k) - \tilde{\nabla} f(x^k) \qquad (\xi^k = \partial f(x^k) - \tilde{\partial} f(x^k)) \tag{2}$$

be random, independent, centered, and have bounded variance:

$$\mathbf{E}\xi^k = 0, \qquad \mathbf{E}\|\xi^k\|^2 \leq \sigma^2. \tag{3}$$

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

$$x^{k+1} = P_Q(x^k - \gamma_k \tilde{\partial} f(x^k)), \qquad \sum_{k=0}^{\infty} \gamma_k^2 < \infty, \quad \sum_{k=0}^{\infty} \gamma_k = \infty, \tag{4}$$

when (2) and (3) are satisfied we have $x^k \to x^*$ a.s., where x^* is some minimum point of $f(x)$ on Q. If $f(x)$ is strongly convex, then $\mathbf{E}\|x^k - x^*\|^2 \to 0$ (the condition $\sum_{k=0}^{\infty} \gamma_k^2 < \infty$ here can be changed to $\gamma_k \to 0$), whereas if $\gamma_k = \gamma/k$ and γ is sufficiently great, then $\mathbf{E}\|x^k - x^*\|^2 = O(1/k)$. □

For sharp minimum problems there is apparently no need to make γ_k tend to zero. One may assume that for a correct adjustment of the step size the 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:

$$\begin{aligned} \bar{x}^k &= \operatorname*{argmin}_{x \in Q} (\tilde{\nabla} f(x^k), x), \\ x^{k+1} &= x^k + \gamma_k(\bar{x}^k - x^k), \qquad \gamma_k \to 0, \ \sum_{k=0}^{\infty} \gamma_k = \infty. \end{aligned} \tag{5}$$

However, such a method does not generally converge. For example, assume we are seeking the minimum of a smooth function $f(x)$, $x \in \mathbf{R}^1$, 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 $\mathbf{E}(\bar{x}^k - x^k) = (\beta - \alpha)/2 \neq 0$ 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:

$$\begin{aligned} \bar{x}^k &= \operatorname*{argmin}_{z \in Q} (y^k, x) , \\ y^k &= y^{k-1} + \mu_k(\tilde{\nabla} f(x^k) - y^{k-1}) , \qquad \mu_k \to 0, \ \sum_{k=0}^{\infty} \mu_k = \infty , \\ x^{k+1} &= x^k + \gamma_k(\bar{x}^k - x^k) , \qquad \gamma_k \to 0, \ \sum_{k=0}^{\infty} \gamma_k = \infty . \end{aligned} \tag{6}$$

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

$$\begin{aligned} &\min f(x), \qquad x \in \mathbf{R}^n , \\ &g_i(x) = 0 , \qquad i = 1,\ldots,m , \end{aligned} \tag{A}$$

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 $Q = \{x: g_i(x) = 0, i = 1,\ldots,m\}$. 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 $f(x^*) \le f(x)$ 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 $y_0^*, y_1^*, \ldots, y_m^*$, 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 $\nabla g_i(x^*)$, $i = 1,...,m$, are linearly independent.

THEOREM 2 (The rule of Lagrange multipliers). If x^* is a regular minimum point, then we can find $y_1^*, ..., y_m^*$ such that

$$\nabla f(x^*) + \sum_{i=1}^{m} y_i^* \nabla g_i(x^*) = 0 . \qquad (2)$$

The $y_1^*, ..., y_m^*$ 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 $\Sigma_{i=1}^m y_i^* \nabla g_i(x^*) = 0$, not all of the $y_1^*, ..., y_m^*$ 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: $\Sigma_{i=1}^m \mu_i \nabla g_i(x^*) = 0$, $\Sigma_{i=1}^m \mu_i^2 \neq 0$, then equality (1) holds for $y_0^* = 0$, $y_i^* = \mu_i$, $i = 1,...,m$. 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

$$L(x, y) = f(x) + (y, g(x)) = f(x) + \sum_{i=1}^{m} y_i g_i(x) \qquad (3)$$

defined on $\mathbf{R}^n \times \mathbf{R}^m$. Here and below we use the vector notation $y = (y_1,...,y_m)$, $g(x) = (g_1(x), ..., g_m(x))$. Then the rule of Lagrange multipliers is:

$$L_x'(x^*, y^*) = 0 , \quad L_y'(x^*, y^*) = 0 , \qquad (4)$$

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

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 $Q = \{x \in \mathbf{R}^n: g_i(x) = 0, i = 1,...,m\}$, where the $g_i(x)$ are continuously differentiable in a neighborhood of $x^* \in Q$, and $\nabla g_i(x^*)$, $i = 1,...,m$, are linearly independent. Then

$$S_Q(x^*) = \{s \in \mathbf{R}^n: (s, \nabla g_i(x^*)) = 0, i = 1,...,m\}, \qquad (5)$$

i.e., the tangent vectors to Q at x^* form a subspace orthogonal to the vectors $\nabla g_1(x^*)$, ..., $\nabla g_m(x^*)$. □

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 $(c, x) \equiv 0$ for $x \in L$.

PROOF. The set $L_1 = \{x \in \mathbf{R}^n: x = A^T y, y \in \mathbf{R}^m\}$ 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 $c \in L_1$. □

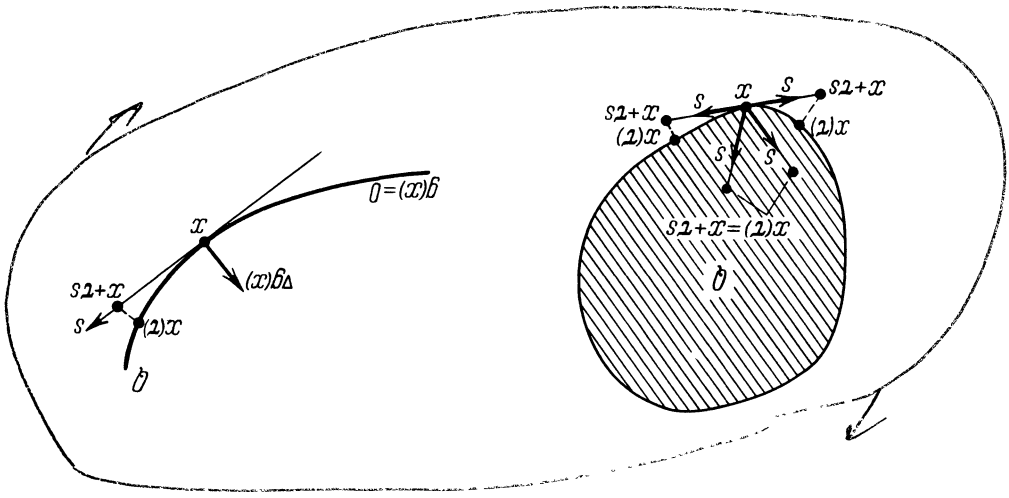


Fig. 33 Tangent vectors.

PROOF of Theorem 2. Let s be a tangent vector to the set $Q = \{x: g_i(x) = 0, i = 1,...,m\}$ at x^*. Then we can find $x(\tau)$ such that $g_i(x(\tau)) = 0$, $i = 1,...,m$, $\|x^* + \tau s - x(\tau)\| = o(\tau)$. 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 τ, then $(\nabla f(x^*), s) \geq 0$. By Lyusternik's theorem, $(s, \nabla g_i(x^*)) = 0$, $i = 1,...,m$. Using Lemma 1, we obtain $\nabla f(x^*) = \sum_{i=1}^m \mu_i \nabla g_i(x^*)$, where the μ_i are some scalars. Setting $y_i^* = -\mu_i$, $i = 1,...,m$, we arrive at (2). □

2. *Proof Based on Elimination of Variables.* If $\nabla g_i(x^*)$ are linearly independent, then the matrix $g'(x^*)$, whose rows are $\nabla g_1(x^*)$, ..., $\nabla g_m(x^*)$, 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 $\partial g_j(x^*)/\partial x_i$, $j = 1,...,m$, $i \in I$, 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

$$[g'_u(u^*, v^*)^T]^{-1} f'_u(u^*, v^*)^T = -y^* . \qquad (6)$$

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). □

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 $U = \{x: \|x-x^*\| \le \varepsilon\}$, 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

$$\min_{x \in U} f_k(x) , \quad f_k(x) = f(x) + \tfrac{1}{2}K\sum_{i=1}^{m} g_i^2(x) + \tfrac{1}{2}\|x-x^*\|^2 , \qquad (7)$$

where K is some parameter. By the continuity of $f_k(x)$, problem (7) has a solution x^k. Therefore

$$f_k(x^k) \le f_k(x^*) ,$$

$$f_k(x) + \tfrac{1}{2}K\sum_{i=1}^{m} g_i^2(x^k) + \tfrac{1}{2}\|x^k - x^*\|^2 \le f(x^*) ,$$

$$\sum_{i=1}^{m} g_i^2(x^k) \le \tfrac{2}{K}\left(f(x^*) - f(x^k) - \tfrac{1}{2}\|x^k - x^*\|^2\right) .$$

The quantity on the right-hand side tends to 0 as $K \to \infty$ (since $\|x^k - x^*\| \le \varepsilon$), therefore $g(x^k) \to 0$. Let a sequence $x^{k_i} \to \bar{x} \in U$. Then $g(\bar{x}) = 0$, $f(\bar{x}) + \|\bar{x}-x^*\|^2/2 \le f(x^*)$; on the other hand, since x^* is a minimum point on Q, then $f(x^*) \le f(\bar{x})$. Hence $\bar{x} = x^*$. Since every limit point for x^k coincides with x^*, one has $x^k \to x^*$ as $K \to \infty$. 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.,

$$\nabla f(x^k) + K\sum_{i=1}^{m} g_i(x^k)\nabla g_i(x^k) + x^k - x^* = 0 . \qquad (8)$$

Let

$$y_0^k = \frac{1}{\sqrt{1 + K^2\Sigma_{i=1}^m g_i^2(x^k)}}, \quad y_i^k = \frac{Kg_i(x^k)}{\sqrt{1 + K^2\Sigma_{i=1}^m g_i^2(x^k)}}, \quad i = 1,...,m .$$

Equality (8) can be written in the form

$$y_0^k\nabla f(x^k) + \sum_{i=1}^{m} y_i^k\nabla g_i(x^k) + (x^k - x^*)y_0^k = 0 . \qquad (9)$$

We have $\Sigma_{i=0}^m (y_i^k)^2 = 1$ for all k, and therefore we can find a sequence $k_j \to \infty$ such that

$$y_i^{k_j} \to y_i^* , \quad i = 0,...,m , \quad \sum_{i=0}^{m} (y_i^*)^2 = 1 .$$

Passing to the limit in (9) yields (1). □

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 ($K \to \infty$) of unconstrained minimization problems that differ by an increasing "penalty" for violating the constraints (the term $(\frac{1}{2})K\Sigma_{i=1}^m g_i^2(x)$ 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 $g_i(x) = (a^i, x) - b_i$, $i = 1,...,m$, 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 Γ, that is the closure of the cone generated by the feasible directions (see the proof of Theorem 3 in Section 7.1).

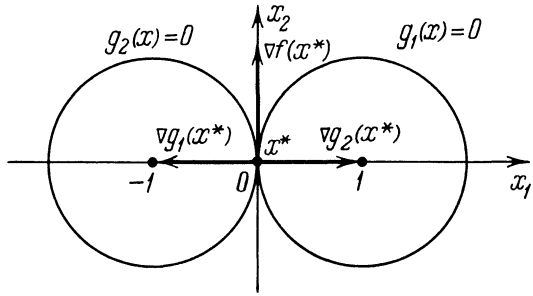


Fig. 34 The problem with a nonregular minimum.

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 $\|x^k - x^*\|^2/2$ 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 $K \to \infty$: $Kg_i(x^k) \to y_i^*/y_0^*$.

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^*, $i = 1,...,m$, be Lagrange multipliers. Then

$$(L''_{xx}(x^*, y^*)s, s) = \left((\nabla^2 f(x^*) + \sum_{i=1}^{m} y_i^* \nabla^2 g_i(x^*))s, s\right) \geq 0 \qquad (10)$$

for all

$$s \in S = \{s: (\nabla g_i(x^*), s) = 0, i = 1,...,m\} .$$

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

$$f(x^*) \le f(x(\tau)) = L(x(\tau), y^*)$$
$$= L(x^*,y^*) + (L'_x(x^*,y^*), x(\tau)-x^*)$$
$$+ (L''_{xx}(x^*,y^*)(x(\tau)-x^*), x(\tau)-x^*)/2 + o(\tau^2)$$
$$= f(x^*) + (\tau^2/2)(L''_{xx}(x^*, y^*)s, s) + o(\tau^2),$$

yielding $(L''_{xx}(x^*, y^*)s, s) \ge 0$. □

The more general necessary extremum condition $L''_{xx}(x^*, y^*) \ge 0$ 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 $x \ne 0$ such that $Cx = 0$. Then the block matrix

$$B = \begin{pmatrix} A & C^T \\ C & 0 \end{pmatrix} \tag{11}$$

of dimension $(m+n) \times (m+n)$ is invertible. □

LEMMA 3. Under the conditions of Lemma 2, there is a $K_0 > 0$, $\alpha > 0$, such that $A + KC^TC \ge \alpha I$ for $K \ge K_0$. □

LEMMA 4. Under the conditions of Lemma 2, for sufficiently large K one has

$$\|(A + KC^TC)^{-1}\| \le \alpha_1/K, \qquad \|C(A + KC^TC)^{-1}\| \le \alpha_2/K,$$
$$\|I - KC(A+KC^TC)^{-1}C^T\| \le \alpha_3/K,$$

where α_i are constants. □

LEMMA 5. Under the conditions of Lemma 2, for the matrix

$$B_k = \begin{pmatrix} A & C^T \\ C & -\frac{1}{K}I \end{pmatrix} \tag{12}$$

for sufficiently large K, B_k^{-1} exists and $\|B_k^{-1}\| \le \gamma$. □

Now let us return to formulating extremum conditions.

THEOREM 5 (sufficient second-order condition). Let $g_i(x^*) = 0$, $i = 1,...,m$, let the functions $f(x)$ and $g_i(x)$ be twice continuously differentiable in a neighborhood of x^*, and let $\nabla g_i(x^*)$, $i = 1,...,m$, be linearly independent. Furthermore, let the necessary minimum condition (4) be satisfied and let

$$(L''_{xx}(x^*, y^*)s, s) > 0 \tag{13}$$

for all s such that $(\nabla g_i(x^*), s) = 0$, $i = 1,...,m$. 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

$$M(x,y,K) = f(x) + (y, g(x)) + \frac{K}{2}\|g(x)\|^2 = L(x,y) + \frac{K}{2}\|g(x)\|^2, \tag{14}$$

where $K > 0$ is some parameter. Then

$$M'_x(x^*, y^*, K) = L'_x(x^*, y^*) = 0,$$

$$M''_{xx}(x^*, y^*) = L''_{xx}(x^*, y^*) + Kg'(x^*)^T g'(x^*).$$

For the matrices $A = L''_{xx}(x^*, y^*)$ and $C = g'(x^*)$ Lemma 3 is applicable. Hence for sufficiently large $K > 0$,

$$M''_{xx}(x^*, y^*, K) > 0. \tag{15}$$

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^*. □

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):

$$M'_x(x^*, y^*, K) = 0, \quad M'_y(x^*, y^*, K) = 0, \tag{16}$$

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

$$\min f(x), x \in \mathbf{R}^2, \quad g(x) = 0, \quad f(x) = x_1^2 - x_2^2, \quad g(x) = x_2,$$

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, $x^* = \operatorname{argmin}_{x \in S} L(x, y^*)$.

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) $\min \Sigma_{i=1}^n x_i^2$, $\Sigma_{i=1}^n x_i = 1$;

(b) $\min \Sigma_{i=1}^n x_i$, $\Sigma_{i=1}^n x_i^2 = 1$;

(c) $\min (Ax, x)$, $\|x\| = 1$;

(d) $\min \|x\|^2$, $(Ax, x) = 1$.

ANSWER: (a) $x_i^* = 1/n$, $i = 1,...,n$; (b) $x_i^* = -1/\sqrt{n}$, $i = 1,...,n$; (c) $x^* = e^1$, the normalized eigenvector corresponding to the ~~largest~~ least eigenvalue of the matrix A; (d) $x^* = \lambda_n^{-1/2} e^n$, e^n being the normalized eigenvector corresponding to the largest eigenvalue λ_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 ~~non~~singular 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

$$f(x) - f(x^*) \geq \ell \|x - x^*\|^2 \tag{17}$$

for $x \in Q$ in a neighborhood of x^*. □

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. □

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

$$\begin{aligned} &\min f(x) , \\ &g_i(x) = \varepsilon_i , \quad i = 1,...,m , \end{aligned} \tag{18}$$

where $\varepsilon = (\varepsilon_1, ..., \varepsilon_m) \in \mathbf{R}^m$ is some vector. Let x_ε denote the solution to this problem (if it exists) and let $\phi(\varepsilon) = f(x_\varepsilon)$. We are interested in the case where $x_\varepsilon \to x^*$ as $\varepsilon \to 0$ (x^* is the solution of (A)), as well as in estimation of the proximity of x_ε to x^* and the behavior of $\phi(\varepsilon)$ for small ε.

THEOREM 8. Let x^* be a nonsingular solution to problem (A). Then for sufficiently small $\|\varepsilon\|$ there exists an x_ε,

$$\|x_\varepsilon - x^*\| = O(\varepsilon) , \quad \nabla\phi(0) = -y^* . \tag{19}$$

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

$$R(z) = 0 . \tag{20}$$

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

$$R'(z^*) = \begin{pmatrix} A & C^T \\ C & 0 \end{pmatrix}, \quad A = L''_{xx}(x^*,y^*), \quad C = L''_{yx}(x^*,y^*) = g'(x^*). \tag{21}$$

It follows from Lemma 2 that $R'(z^*)$ is nonsingular. By Theorem 3 of Section 2.3 the system

$$R(z) = a \tag{22}$$

has a solution z_a for sufficiently small $\|a\|$, and

$$z_a = z^* - [R'(z^*)]^{-1} a + o(a). \tag{23}$$

Take $a = \{0, \varepsilon\}$, $a \in \mathbf{R}^{n+m}$, $\varepsilon \in \mathbf{R}^m$. In this case (22) is equivalent to the system

$$\nabla f(x) + g'(x)^T y = 0, \qquad g(x) = \varepsilon, \tag{24}$$

and it has a solution $z_a = \{x_\varepsilon, y_\varepsilon\}$ for sufficiently small ε. Therefore, the point x_ε (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_ε we have, by (24), the necessary minimum condition in problem (18) with Lagrange multipliers y_ε, 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 $S_\varepsilon = \{s: (\nabla g_i(x_\varepsilon), s) = 0, i = 1,...,m\}$. Thus, at x_ε we have the sufficient second-order extremum condition, i.e., x_ε is a solution of problem (18). It follows from (23) that $\|z_a - z^*\| \le \alpha\|a\|$, α being some constant, and hence $\|x_\varepsilon - x^*\| \le \alpha\|\varepsilon\|$. Finally,

$$\begin{aligned} f(x) \ge f(x) + (y^*, g(x)) &= L(x,y^*) \ge L(x^*,y^*) \\ &= f(x^*) + (y^*, g(x^*)) = f(x^*), \end{aligned}$$

$$\begin{aligned} \phi(\varepsilon) = f(x_\varepsilon) &= f(x^*) + (\nabla f(x^*), x_\varepsilon - x^*) + o(\|x_\varepsilon - x^*\|) \\ &= f(x^*) - (g'(x^*)^T y^*, x_\varepsilon - x^*) + o(\varepsilon) \\ &= f(x^*) - (g(x_\varepsilon) - g(x^*), y^*) + o(\varepsilon) = f(x^*) - (y^*, \varepsilon) + o(\varepsilon). \end{aligned}$$

Therefore, $\nabla\phi(0) = -y^*$. □