Constrained Optimization and Lagrange Multiplier Methods

Dimitri P. Bertsekas

Massachusetts Institute of Technology

WWW site for book information and orders

http://athenasc.com

Athena Scientific, Belmont, Massachusetts

Athena Scientific
Post Office Box 391
Belmont, Mass. 02178-9998
U.S.A.

Email: info@athenasc.com
WWW information and orders: http://athenasc.com

Cover Design: *Ann Gallager*

Originally published by Academic Press, Inc., in 1982

Publisher's Cataloging-in-Publication Data

Bertsekas, Dimitri P.
Constrained Optimization and Lagrange Multiplier Methods
Includes bibliographical references and index
1. Mathematical Optimization. 2. Multipliers (Mathematical Analysis). I. Title.
QA402.5.B46 1996 519.4 96-79307

ISBN 1-886529-04-3

ABOUT THE AUTHOR

Dimitri Bertsekas studied Mechanical and Electrical Engineering at the National Technical University of Athens, Greece, and obtained his Ph.D. in system science from the Massachusetts Institute of Technology. He has held faculty positions with the Engineering-Economic Systems Department, Stanford University, and the Electrical Engineering Department of the University of Illinois, Urbana. Since 1979 he has been teaching at the Electrical Engineering and Computer Science Department of the Massachusetts Institute of Technology (M.I.T.), where he is McAfee Professor of Engineering. In 2019, he joined the School of Computing, Informatics, and Decision Systems Engineering at the Arizona State University, Tempe, AZ, as Fulton Professor of Computational Decision Making.

Professor Bertsekas' teaching and research have spanned several fields, including deterministic optimization, dynamic programming and stochastic control, large-scale and distributed computation, artificial intelligence, and data communication networks. He has authored or coauthored numerous research papers and eighteen books, several of which are currently used as textbooks in MIT classes, including "Dynamic Programming and Optimal Control," "Data Networks," "Introduction to Probability," and "Nonlinear Programming."

Professor Bertsekas was awarded the INFORMS 1997 Prize for Research Excellence in the Interface Between Operations Research and Computer Science for his book "Neuro-Dynamic Programming" (co-authored with John Tsitsiklis), the 2001 AACC John R. Ragazzini Education Award, the 2009 INFORMS Expository Writing Award, the 2014 AACC Richard Bellman Heritage Award, the 2014 INFORMS Khachiyan Prize for Life-Time Accomplishments in Optimization, the 2015 MOS/SIAM George B. Dantzig Prize, and the 2022 IEEE Control Systems Award. In 2018 he shared with his coauthor, John Tsitsiklis, the 2018 INFORMS John von Neumann Theory Prize for the contributions of the research monographs "Parallel and Distributed Computation" and "Neuro-Dynamic Programming." Professor Bertsekas was elected in 2001 to the United States National Academy of Engineering for "pioneering contributions to fundamental research, practice and education of optimization/control theory, and especially its application to data communication networks."

ATHENA SCIENTIFIC
OPTIMIZATION AND COMPUTATION SERIES

1. Rollout, Policy Iteration, and Distributed Reinforcement Learning, by Dimitri P. Bertsekas, 2020, ISBN 978-1-886529-07-6, 480 pages
2. Reinforcement Learning and Optimal Control, by Dimitri P. Bertsekas, 2019, ISBN 978-1-886529-39-7, 388 pages
3. Abstract Dynamic Programming, 2nd Edition, by Dimitri P. Bertsekas, 2018, ISBN 978-1-886529-46-5, 360 pages
4. Dynamic Programming and Optimal Control, Two-Volume Set, by Dimitri P. Bertsekas, 2017, ISBN 1-886529-08-6, 1270 pages
5. Nonlinear Programming, 3rd Edition, by Dimitri P. Bertsekas, 2016, ISBN 1-886529-05-1, 880 pages
6. Convex Optimization Algorithms, by Dimitri P. Bertsekas, 2015, ISBN 978-1-886529-28-1, 576 pages
7. Convex Optimization Theory, by Dimitri P. Bertsekas, 2009, ISBN 978-1-886529-31-1, 256 pages
8. Introduction to Probability, 2nd Edition, by Dimitri P. Bertsekas and John N. Tsitsiklis, 2008, ISBN 978-1-886529-23-6, 544 pages
9. Convex Analysis and Optimization, by Dimitri P. Bertsekas, Angelia Nedić, and Asuman E. Ozdaglar, 2003, ISBN 1-886529-45-0, 560 pages
10. Network Optimization: Continuous and Discrete Models, by Dimitri P. Bertsekas, 1998, ISBN 1-886529-02-7, 608 pages
11. Network Flows and Monotropic Optimization, by R. Tyrrell Rockafellar, 1998, ISBN 1-886529-06-X, 634 pages
12. Introduction to Linear Optimization, by Dimitris Bertsimas and John N. Tsitsiklis, 1997, ISBN 1-886529-19-1, 608 pages
13. Parallel and Distributed Computation: Numerical Methods, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1997, ISBN 1-886529-01-9, 718 pages
14. Neuro-Dynamic Programming, by Dimitri P. Bertsekas and John N. Tsitsiklis, 1996, ISBN 1-886529-10-8, 512 pages
15. Constrained Optimization and Lagrange Multiplier Methods, by Dimitri P. Bertsekas, 1996, ISBN 1-886529-04-3, 410 pages
16. Stochastic Optimal Control: The Discrete-Time Case, by Dimitri P. Bertsekas and Steven E. Shreve, 1996, ISBN 1-886529-03-5, 330 pages

To Teli and Taki

Contents

Preface

The area of Lagrange multiplier methods for constrained minimization has undergone a radical transformation starting with the introduction of augmented Lagrangian functions and methods of multipliers in 1968 by Hestenes and Powell. The initial success of these methods in computational practice motivated further efforts aimed at understanding and improving their properties. At the same time their discovery provided impetus and a new perspective for reexamination of Lagrange multiplier methods proposed and nearly abandoned several years earlier. These efforts, aided by fresh ideas based on exact penalty functions, have resulted in a variety of interesting methods utilizing Lagrange multiplier iterations and competing with each other for solution of different classes of problems.

This monograph is the outgrowth of the author's research involvement in the area of Lagrange multiplier methods over a nine-year period beginning in early 1972. It is aimed primarily toward researchers and practitioners of mathematical programming algorithms, with a solid background in introductory linear algebra and real analysis.

Considerable emphasis is placed on the method of multipliers which, together with its many variations, may be viewed as a primary subject of the monograph. Chapters 2, 3, and 5 are devoted to this method. A large portion of Chapter 1 is devoted to unconstrained minimization algorithms on which

the method relies. The developments on methods of multipliers serve as a good introduction to other Lagrange multiplier methods examined in Chapter 4.

Several results and algorithms were developed as the monograph was being written and have not as yet been published in journals. These include the algorithm for minimization subject to simple constraints (Section 1.5), the improved convergence and rate-of-convergence results of Chapter 2, the first stepsize rule of Section 2.3.1, the unification of the exact penalty methods of DiPillo and Grippo, and Fletcher, and their relationship with Newton's method (Section 4.3), the globally convergent Newton and quasi-Newton methods based on differentiable exact penalty functions (Section 4.5.2), and the methodology for solving large-scale separable integer programming problems of Section 5.6.

The line of development of the monograph is based on the author's conviction that solving practical nonlinear optimization problems efficiently (or at all) is typically a challenging undertaking and can be accomplished only through a thorough understanding of the underlying theory. This is true even if a polished packaged optimization program is used, but more so when the problem is large enough or important enough to warrant the development of a specialized algorithm. Furthermore, it is quite common in practice that methods are modified, combined, and extended in order to construct an algorithm that matches best the features of the particular problem at hand, and such modifications require a full understanding of the theoretical foundations of the method utilized. For these reasons, we place primary emphasis on the principles underlying various methods and the analysis of their convergence and rate-of-convergence properties. We also provide extensive guidance on the merits of various types of methods but, with a few exceptions, do not provide any algorithms that are specified to the last level of detail.

The monograph is based on the collective works of many researchers as well as my own. Of those people whose work had a substantial influence on my thinking and contributed in an important way to the monograph I would like to mention J. D. Buys, G. DiPillo, L. Dixon, R. Fletcher, T. Glad, L. Grippo, M. Hestenes, D. Luenberger, O. Mangasarian, D. Q. Mayne, E. Polak, B. T. Poljak, M. J. D. Powell, B. Pschenichny, R. T. Rockafellar, and R. Tapia. My research on methods of multipliers began at Stanford University. My interaction there with Daniel Gabay, Barry Kort, and David Luenberger had a lasting influence on my subsequent work on the subject. The material of Chapter 5 in particular is largely based on the results of my direct collaboration with Barry Kort. The material of Sec-

tion 5.6 is based on work on electric power system scheduling at Alphatech, Inc. where I collaborated with Greg Lauer, Tom Posbergh, and Nils R. Sandell, Jr.

Finally, I wish to acknowledge gratefully the research support of the National Science Foundation, and the expert typing of Margaret Flaherty, Leni Gross, and Rosalie J. Bialy.

Chapter 1

Introduction

1.1 General Remarks

Two classical nonlinear programming problems are the equality constrained problem

(ECP) minimize $f(x)$

subject to $h(x) = 0$

and its inequality constrained version

(ICP) minimize $f(x)$

subject to $g(x) \leq 0,$

where $f: R^n \to R$, $h: R^n \to R^m$, $g: R^n \to R^r$ are given functions. Computational methods for solving these problems became the subject of intensive investigation during the late fifties and early sixties. We discuss three of the approaches that were pursued.

The first approach was based on the idea of iterative descent within the confines of the constraint set. Given a feasible point x_k, a direction d_k was chosen satisfying the descent condition $\nabla f(x_k)'d_k < 0$ and the condition

$x_k + \alpha d_k$: feasible for all α positive and sufficiently small. A search along the line $\{x_k + \alpha d_k | \alpha > 0\}$ produced a new feasible point $x_{k+1} = x_k + \alpha_k d_k$ satisfying $f(x_{k+1}) < f(x_k)$. This led to various classes of *feasible direction methods* with which the names of Frank–Wolfe, Zoutendijk, Rosen, Goldstein, and Levitin–Poljak are commonly associated. These methods, together with their more sophisticated versions, enjoyed considerable success and still continue to be very popular for problems with linear constraints. On the other hand, feasible direction methods by their very nature were unable to handle problems with nonlinear equality constraints, and some of them were inapplicable or otherwise not well suited for handling nonlinear inequality constraints as well. A number of modifications were proposed for treating nonlinear equality constraints, but these involved considerable complexity and detracted substantially from the appeal of the descent idea.

A second approach was based on the possibility of solving the system of equations and (possibly) inequalities which constitute necessary conditions for optimality for the optimization problem. For (ECP), these conditions are

$$\nabla_x L(x, \lambda) = \nabla f(x) + \nabla h(x)\lambda = 0, \tag{1a}$$

$$\nabla_\lambda L(x, \lambda) = h(x) = 0, \tag{1b}$$

where L is the (ordinary) Lagrangian function

$$L(x, \lambda) = f(x) + \lambda' h(x).$$

A distinguishing feature of this approach is that the Lagrange multiplier λ is treated on an equal basis with the vector x. Iterations are carried out simultaneously on x and λ, by contrast with the descent approach where only x is iterated upon and the Lagrange multiplier plays no direct role. For this reason algorithms of this type are sometimes called *Lagrangian methods*. Several methods of this type were considered in Arrow *et al.* (1958). In addition to Newton's method for solving system (1), a gradient method was also proposed under the condition that the *local convexity assumption*

$$\nabla^2_{xx} L(x^*, \lambda^*) > 0 \tag{2}$$

holds at a solution (x^*, λ^*). It was noted, however, by Arrow and Solow (1958) that if the local convexity assumption did not hold, then (ECP) could be replaced by the equivalent problem

$$\begin{aligned} &\text{minimize} \quad f(x) + \tfrac{1}{2}c|h(x)|^2 \\ &\text{subject to} \quad h(x) = 0, \end{aligned} \tag{3}$$

where c is a scalar and $|\cdot|$ denotes Euclidean norm. If c is taken sufficiently large, then the local convexity condition can be shown to hold for problem (3) under fairly mild conditions. The idea of focusing attention on the necessary

conditions rather than the original problem also attracted considerable attention in optimal control where the necessary conditions can often be formulated as a two-point boundary value problem. However, it quickly became evident that the approach had some fundamental limitations, mainly the lack of a good mechanism to enforce convergence when far from a solution, and the difficulty of some of the methods to distinguish between local minima and local maxima.

A third approach was based on elimination of constraints through the use of *penalty functions*. For example the quadratic penalty function method (Fiacco and McCormick, 1968) for (ECP) consists of sequential unconstrained minimization of the form

$$\begin{aligned} &\text{minimize} \quad f(x) + \tfrac{1}{2}c_k|h(x)|^2 \\ &\text{subject to} \quad x \in R^n, \end{aligned} \tag{4}$$

where $\{c_k\}$ is a positive scalar sequence with $c_k < c_{k+1}$ for all k and $c_k \to \infty$. The sequential minimization process yields

$$\lim_{c_k \to \infty} \inf_{x \in R^n} \{f(x) + \tfrac{1}{2}c_k|h(x)|^2\}. \tag{5}$$

On the other hand, the optimal value of (ECP) can be written as

$$\inf_{x \in R^n} \lim_{c_k \to \infty} \{f(x) + \tfrac{1}{2}c_k|h(x)|^2\}, \tag{6}$$

and hence the success of the penalty method hinges on the equality of the expressions (5) and (6), i.e., the validity of interchanging "lim" and "inf." This interchange is indeed valid under mild assumptions (basically continuity of f and h—see Chapter 2). Lagrange multipliers play no direct role in this method but it can be shown under rather mild assumptions that the sequence $\{c_k h(x_k)\}$, where x_k solves problem (4), converges to a Lagrange multiplier of the problem. Despite their considerable disadvantages [mainly slow convergence and ill-conditioning when solving problem (4) for large values of c_k], penalty methods were widely accepted in practice. The reasons can be traced to the simplicity of the approach, its ability to handle nonlinear constraints, as well as the availability of very powerful unconstrained minimization methods for solving problem (4).

The main idea of the descent approach also made its appearance in a dual context whereby an ascent method is used to maximize the *dual functional* for (ECP) given by

$$d(\lambda) = \inf_x \{f(x) + \lambda' h(x)\} = \inf_x L(x, \lambda).$$

In the simplest such method one minimizes $L(\cdot, \lambda_k)$ (perhaps in a local sense) over x for a sequence of multiplier vectors $\{\lambda_k\}$. This sequence is generated by

$$\lambda_{k+1} = \lambda_k + \alpha h(x_k), \tag{7}$$

where x_k is a minimizing point of $L(\cdot, \lambda_k)$ and α is a stepsize scalar parameter. It is possible to show under the appropriate assumptions (see Section 2.6) that $h(x_k) = \nabla d(\lambda_k)$, so (7) is actually a steepest ascent iteration for maximizing the dual functional d. Such methods have been called *primal–dual methods*. Actually the dual functional and the method itself make sense only under fairly restrictive conditions including either the local convexity assumption (2) or other types of convexity conditions. The method is also often hampered by slow convergence. Furthermore in many cases it is difficult to know a priori an appropriate range for the stepsize α. For this reason primal–dual methods of the type just described initially found application only in the limited class of convex or locally convex problems where minimization of $L(\cdot, \lambda_k)$ can be carried out very efficiently due to special structure involving, for example, separable objective and constraint functions (Everett, 1963).

Starting around 1968, a number of researchers have proposed a new class of methods, called *methods of multipliers*, in which the penalty idea is merged with the primal–dual and Lagrangian philosophy. In the original method of multipliers, proposed by Hestenes (1969) and Powell (1969), the quadratic penalty term is added not to the objective function f of (ECP) but rather to the Lagrangian function $L = f + \lambda' h$ thus forming the *augmented Lagrangian* function

$$L_c(x, \lambda) = f(x) + \lambda' h(x) + \tfrac{1}{2} c |h(x)|^2. \tag{8}$$

A sequence of minimizations of the form

$$\begin{aligned} &\text{minimize} \quad L_{c_k}(x, \lambda_k) \\ &\text{subject to} \quad x \in R^n \end{aligned} \tag{9}$$

is performed where $\{c_k\}$ is a sequence of positive penalty parameters. The multiplier sequence $\{\lambda_k\}$ is generated by the iteration

$$\lambda_{k+1} = \lambda_k + c_k h(x_k), \tag{10}$$

where x_k is a solution of problem (9). The initial vector λ_0 is selected a priori, and the sequence $\{c_k\}$ may be either preselected or generated during the computation according to some scheme.

One may view the method just described within the context of penalty function methods. If $c_k \to \infty$ and the generated sequence $\{\lambda_k\}$ turns out to

be bounded, then the method is guaranteed to yield in the limit the optimal value of (ECP), provided sufficient assumptions are satisfied which guarantee the validity of interchange of "lim" and "inf" in the expression

$$\lim_{c_k \to \infty} \inf_x \{f(x) + \lambda_k' h(x) + \tfrac{1}{2} c_k |h(x)|^2\},$$

similarly as for the penalty method considered earlier.

Another point of view (see Chapter 2) is based on the fact that iteration (10) is a steepest ascent iteration for maximizing the dual functional

$$d_{c_k}(\lambda) = \inf_x \{f(x) + \lambda' h(x) + \tfrac{1}{2} c_k |h(x)|^2\},$$

which corresponds to the problem

$$\begin{aligned} &\text{minimize} \quad f(x) + \tfrac{1}{2} c_k |h(x)|^2 \\ &\text{subject to} \quad h(x) = 0. \end{aligned}$$

As noted earlier, if c_k is sufficiently large, this problem has locally convex structure, so the primal–dual viewpoint is applicable.

It turns out that, by combining features of the penalty and the primal–dual approach, the method of multipliers actually moderates the disadvantages of both. As we shall see in the next chapter, convergence in the method of multipliers can usually be attained *without the need to increase* c_k *to infinity* thereby alleviating the ill-conditioning problem that plagues the penalty method. In addition *the multiplier iteration* (10) *tends to converge to a Lagrange multiplier vector much faster than iteration* (7) *of the primal–dual method, or the sequence* $\{c_k h(x_k)\}$ *in the penalty method.* Because of these attractive characteristics, the method of multipliers and its subsequently developed variations have emerged as a very important class of constrained minimization methods. A great deal of research has been directed toward their analysis and understanding. Furthermore their discovery provided impetus for reexamination of Lagrangian methods proposed and nearly abandoned many years ago. These efforts aided by fresh ideas based on penalty functions and duality have resulted in a variety of interesting methods utilizing Lagrange multiplier iterations and competing with each other for solution of different classes of problems.

The purpose of this monograph is to provide a rather thorough analysis of these Lagrange multiplier methods starting with the quadratic method of multipliers for (ECP) just described. This method is the subject of Chapter 2. In Chapter 3, the method is extended to handle problems with both equality and inequality constraints. In addition the Lagrange multiplier approach is utilized to construct algorithms for solution of nondifferentiable and minimax problems. In Chapter 4, we consider a variety of Lagrangian methods and

analyze their local and global convergence properties. Finally, in Chapter 5, we explore the possibility of using a penalty function other than quadratic, and we analyze multiplier methods as applied to convex programming problems.

1.2 Notation and Mathematical Background

The purpose of this section is to provide a selective list of mathematical definitions, notations, and results that will be frequently used. For detailed expositions, the reader should consult texts on linear algebra and real analysis.

Algebraic Notions

We denote by R the real line and by R^n the space of all n-dimensional vectors. Intervals of real numbers or extended real numbers are denoted as usual by bracket–parentheses notation. For example for $a \in R$ or $a = -\infty$ and $b \in R$ or $b = +\infty$ we write $(a, b] = \{x \mid a < x \leq b\}$. Given any subset $S \subset R$ which is bounded above (below), we denote by $\sup S$ ($\inf S$) the least upper bound (greatest lower bound) of S. If S is unbounded above (below) we write $\sup S = \infty$ ($\inf S = -\infty$). In our notation, *every vector is considered to be a column vector*. The transpose of an $m \times n$ matrix A is denoted A'. A vector $x \in R^n$ will be treated as an $n \times 1$ matrix, and thus x' denotes a $1 \times n$ matrix or row vector. If $x_1, \ldots, x_n$ are the coordinates of a vector $x \in R^n$, we write $x = (x_1, x_2, \ldots, x_n)$. We also write

$$x \geq 0 \qquad \text{if} \quad x_i \geq 0, \quad i = 1, \ldots, n,$$

$$x \leq 0 \qquad \text{if} \quad x_i \leq 0, \quad i = 1, \ldots, n.$$

A symmetric $n \times n$ matrix A will be said to be *positive semidefinite* if $x'Ax \geq 0$ for all $x \in R^n$. In this case we write

$$A \geq 0.$$

We say that A is *positive definite* if $x'Ax > 0$ for all $x \neq 0$, and write

$$A > 0.$$

When we say that A is positive (semi)definite we implicitly assume that it is symmetric. A symmetric $n \times n$ matrix A has n real eigenvalues $\gamma_1, \gamma_2, \ldots, \gamma_n$ and n nonzero real eigenvectors $e_1, e_2, \ldots, e_n$ which are mutually orthogonal. It can be shown that

$$\gamma x'x \leq x'Ax \leq \Gamma x'x \qquad \forall\, x \in R^n, \tag{1}$$

where

$$\gamma = \min\{\gamma_1, \ldots, \gamma_n\}, \qquad \Gamma = \max\{\gamma_1, \ldots, \gamma_n\}.$$

For x equal to the eigenvector corresponding to Γ (γ), the inequality on the right (left) in (1) becomes equality. It follows that $A > 0$ ($A \geq 0$), if and only if the eigenvalues of A are positive (nonnegative).

If A is positive definite, there exists a unique positive definite matrix the square of which equals A. This is the matrix that has the same eigenvectors as A and has as eigenvalues the square roots of the eigenvalues of A. We denote this matrix by $A^{1/2}$.

Let A and B be square matrices and C be a matrix of appropriate dimension. The very useful equation

$$(A + CBC')^{-1} = A^{-1} - A^{-1}C(B^{-1} + C'A^{-1}C)^{-1}C'A^{-1}$$

holds provided all the inverses appearing above exist. The equation can be verified by multiplying the right-hand side by $(A + CBC')$ and showing that the product is the identity.

Consider a partitioned square matrix M of the form

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}.$$

There holds

$$M^{-1} = \begin{bmatrix} Q & -QBD^{-1} \\ -D^{-1}CQ & D^{-1} + D^{-1}CQBD^{-1} \end{bmatrix},$$

where

$$Q = (A - BD^{-1}C)^{-1},$$

provided all the inverses appearing above exist. The proof is obtained by multiplying M with the expression for M^{-1} given above and verifying that the product yields the identity matrix.

Topological Notions

We shall use throughout the standard Euclidean norm in R^n denoted $|\cdot|$; i.e., for a vector $x \in R^n$, we write

$$|x| = \sqrt{x'x}.$$

The Euclidean norm of an $m \times n$ matrix A will be denoted also $|\cdot|$. It is given by

$$|A| = \max_{x \neq 0} \frac{|Ax|}{|x|} = \max_{x \neq 0} \frac{\sqrt{x'A'Ax}}{\sqrt{x'x}}.$$

In view of (1), we have

$$|A| = \sqrt{\max \text{ eigenvalue}(A'A)}.$$

If A is symmetric, then if $\lambda_1, \ldots, \lambda_n$ are its (real) eigenvalues, the eigenvalues of A^2 are $\lambda_1^2, \ldots, \lambda_n^2$, and we obtain

$$|A| = \max\{|\lambda_1|, \ldots, |\lambda_n|\}.$$

A sequence of vectors $x_0, x_1, \ldots, x_k, \ldots$, in R^n, denoted $\{x_k\}$, is said to converge to a limit vector x if $|x_k - x| \to 0$ as $k \to \infty$ (that is, if given $\varepsilon > 0$, there is an N such that for all $k \geq N$ we have $|x_k - x| < \varepsilon$). If $\{x_k\}$ converges to x we write $x_k \to x$ or $\lim_{k\to\infty} x_k = x$. Similarly for a sequence of $m \times n$ matrices $\{A_k\}$, we write $A_k \to A$ or $\lim_{k\to\infty} A_k = A$ if $|A_k - A| \to 0$ as $k \to \infty$. Convergence of both vector and matrix sequences is equivalent to convergence of each of the sequences of their coordinates or elements.

Given a sequence $\{x_k\}$, the subsequence $\{x_k | k \in K\}$ corresponding to an infinite index set K is denoted $\{x_k\}_K$. A vector x is said to be a *limit point* of a sequence $\{x_k\}$ if there is a subsequence $\{x_k\}_K$ which converges to x.

A sequence of real numbers $\{r_k\}$ which is monotonically nondecreasing (nonincreasing), i.e., satisfies $r_k \leq r_{k+1}$ ($r_k \geq r_{k+1}$) for all k, must either converge to a real number or be unbounded above (below) in which case we write $\lim_{k\to\infty} r_k = +\infty$ ($\lim_{k\to\infty} r_k = -\infty$). Given any bounded sequence of real numbers $\{r_k\}$, we may consider the sequence $\{s_k\}$ where $s_k = \sup\{r_i | i \geq k\}$. Since this sequence is monotonically nonincreasing and bounded, it must have a limit called the *limit superior* of $\{r_k\}$ and denoted by $\limsup_{k\to\infty} r_k$. We define similarly the *limit inferior* of $\{r_k\}$ and denote it by $\liminf_{k\to\infty} r_k$. If $\{r_k\}$ is unbounded above, we write $\limsup_{k\to\infty} r_k = +\infty$, and if it is unbounded below, we write $\liminf_{k\to\infty} r_k = -\infty$.

Open, Closed, and Compact Sets

For a vector $x \in R^n$ and a scalar $\varepsilon > 0$, we denote the open sphere centered at x with radius $\varepsilon > 0$ by $S(x; \varepsilon)$; i.e.,

$$S(x; \varepsilon) = \{z \,|\, |z - x| < \varepsilon\}. \tag{2}$$

For a subset $X \subset R^n$ and a scalar $\varepsilon > 0$, we write by extension of the preceding notation

$$S(X; \varepsilon) = \{z \,|\, |z - x| < \varepsilon \text{ for some } x \in X\}. \tag{3}$$

A subset S of R^n is said to be *open*, if for every vector $x \in S$ one can find an $\varepsilon > 0$ such that $S(x; \varepsilon) \subset S$. If S is open and $x \in S$, then S is said to be a *neighborhood* of x. The *interior* of a set $S \subset R^n$ is the set of all $x \in S$ for which there exists $\varepsilon > 0$ such that $S(x; \varepsilon) \subset S$. A set S is *closed* if and only if its

complement in R^n is open. Equivalently S is closed if and only if every convergent sequence $\{x_k\}$ with elements in S converges to a point which also belongs to S. A subset S of R^n is said to be *compact* if and only if it is both closed and bounded (i.e., it is closed and for some $M > 0$ we have $|x| \leq M$ for all $x \in S$). A set S is compact if and only if every sequence $\{x_k\}$ with elements in S has at least one limit point which belongs to S. Another important fact is that if $S_0, S_1, \ldots, S_k, \ldots$ is a sequence of nonempty compact sets in R^n such that $S_k \supset S_{k+1}$ for all k then the intersection $\bigcap_{k=0}^{\infty} S_k$ is a nonempty and compact set.

Continuous Functions

A function f mapping a set $S_1 \subset R^n$ into a set $S_2 \subset R^m$ is denoted by $f: S_1 \to S_2$. The function f is said to be *continuous* at $x \in S_1$ if $f(x_k) \to f(x)$ whenever $x_k \to x$. Equivalently f is continuous at x if given $\varepsilon > 0$ there is a $\delta > 0$ such that $|y - x| < \delta$ and $y \in S_1$ implies $|f(y) - f(x)| < \varepsilon$. The function f is said to be continuous over S_1 (or simply continuous) if it is continuous at every point $x \in S_1$. If S_1, S_2, and S_3 are sets and $f_1: S_1 \to S_2$ and $f_2: S_2 \to S_3$ are functions, the function $f_2 \cdot f_1 : S_1 \to S_3$ defined by $(f_2 \cdot f_1)(x) = f_2[f_1(x)]$ is called the *composition* of f_1 and f_2. If $f_1 : R^n \to R^m$ and $f_2: R^m \to R^p$ are continuous, then $f_2 \cdot f_1$ is also continuous.

Differentiable Functions

A real-valued function $f: X \to R$ where $X \subset R^n$ is an open set is said to be *continuously differentiable* if the partial derivatives $\partial f(x)/\partial x_1, \ldots, \partial f(x)/\partial x_n$ exist for each $x \in X$ and are continuous functions of x over X. In this case we write $f \in C^1$ over X. More generally we write $f \in C^p$ over X for a function $f: X \to R$, where $X \subset R^n$ is an open set if all partial derivatives of order p exist and are continuous as functions of x over X. If $f \in C^p$ over R^n, we simply write $f \in C^p$. If $f \in C^1$ on X, the *gradient* of f at a point $x \in X$ is defined to be the column vector

$$\nabla f(x) = \begin{bmatrix} \dfrac{\partial f(x)}{\partial x_1} \\ \vdots \\ \dfrac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

If $f \in C^2$ over X, the *Hessian* of f at x is defined to be the symmetric $n \times n$ matrix having $\partial^2 f(x)/\partial x_i \, \partial x_j$ as the ijth element

$$\nabla^2 f(x) = \left[\frac{\partial^2 f(x)}{\partial x_i \, \partial x_j}\right].$$

If $f: X \to R^m$ where $X \subset R^n$, then f will be alternatively represented by the column vector of its component functions $f_1, f_2, \ldots, f_m$

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix}.$$

If X is open, we write $f \in C^p$ on X if $f_1 \in C^p, f_2 \in C^p, \ldots, f_m \in C^p$ on X. We shall use the notation

$$\nabla f(x) = [\nabla f_1(x) \cdots \nabla f_m(x)].$$

Thus, the $n \times m$ matrix ∇f has as columns the gradients $\nabla f_1(x), \ldots, \nabla f_m(x)$ and is the transpose of the Jacobian matrix of the function f.

On occasion we shall need to consider gradients of functions with respect to some of the variables only. The notation will be as follows:

If $f: R^{n+r} \to R$ is a real-valued function of (x, y) where $x = (x_1, \ldots, x_n) \in R^n, y = (y_1, \ldots, y_r) \in R^r$, we write

$$\nabla_x f(x, y) = \begin{bmatrix} \dfrac{\partial f(x, y)}{\partial x_1} \\ \vdots \\ \dfrac{\partial f(x, y)}{\partial x_n} \end{bmatrix}, \qquad \nabla_y f(x, y) = \begin{bmatrix} \dfrac{\partial f(x, y)}{\partial y_1} \\ \vdots \\ \dfrac{\partial f(x, y)}{\partial y_r} \end{bmatrix},$$

$$\nabla_{xx} f(x, y) = \left[\frac{\partial f(x, y)}{\partial x_i \, \partial x_j}\right], \qquad \nabla_{xy} f(x, y) = \left[\frac{\partial f(x, y)}{\partial x_i \, \partial y_j}\right],$$

$$\nabla_{yy} f(x, y) = \left[\frac{\partial f(x, y)}{\partial y_i \, \partial y_j}\right].$$

If $f: R^{n+r} \to R^m, f = (f_1, f_2, \ldots, f_m)$, we write

$$\nabla_x f(x, y) = [\nabla_x f_1(x, y) \cdots \nabla_x f_m(x, y)],$$
$$\nabla_y f(x, y) = [\nabla_y f_1(x, y) \cdots \nabla_y f_m(x, y)].$$

For $h: R^r \to R^m$ and $g: R^n \to R^r$, consider the function $f: R^n \to R^m$ defined by

$$f(x) = h[g(x)].$$

Then if $h \in C^p$ and $g \in C^p$, we also have $f \in C^p$. The *chain rule* of differentiation is stated in terms of our notation as

$$\nabla f(x) = \nabla g(x) \nabla h[g(x)].$$

Mean Value Theorems and Taylor Series Expansions

Let $f: X \to R$, and $f \in C^1$ over the open set $X \subset R^n$. Assume that X contains the line segment connecting two points $x, y \in X$. The *mean value theorem* states that there exists a scalar α with $0 < \alpha < 1$ such that

$$f(y) = f(x) + \nabla f[x + \alpha(y - x)]'(y - x).$$

If in addition $f \in C^2$, then there exists a scalar α with $0 < \alpha < 1$ such that

$$f(y) = f(x) + \nabla f(x)'(y - x) + \tfrac{1}{2}(y - x)'\nabla^2 f[x + \alpha(y - x)](y - x).$$

Let $f: X \to R^m$ and $f \in C^1$ on the open set $X \subset R^n$. Assume that X contains the line segment connecting two points $x, y \in X$. The *first-order Taylor series expansion* of f around x is given by the equation

$$f(y) = f(x) + \int_0^1 \nabla f[x + \alpha(y - x)]'(y - x)\, d\alpha.$$

If in addition $f \in C^2$ on X, then we have the *second-order Taylor series expansion*

$$f(y) = f(x) + \nabla f(x)'(y - x) + \int_0^1 \left(\int_0^\xi (y - x)'\nabla^2 f[x + \alpha(y - x)](y - x)\, d\alpha \right) d\xi.$$

Implicit Function Theorems

Consider a system of n equations in $m + n$ variables

$$h(x, y) = 0,$$

where $h: R^{m+n} \to R^n$, $x \in R^m$, and $y \in R^n$. Implicit function theorems address the question whether one may solve the system of equations for the vector y in terms of the vector x, i.e., whether there exists a function ϕ, called the *implicit function*, such that $h[x, \phi(x)] = 0$. The following classical implicit function theorem asserts that this is possible in a local sense, i.e., in a neighborhood of a solution $(\bar{x}, \bar{y})$, provided the gradient matrix of h with respect to y is nonsingular.

Implicit Function Theorem 1: Let S be an open subset of R^{m+n}, and $h: S \to R^n$ be a function such that for some $p \geq 0$, $h \in C^p$ over S, and assume that $\nabla_y h(x, y)$ exists and is continuous on S. Let $(\bar{x}, \bar{y}) \in S$ be a vector such that $h(\bar{x}, \bar{y}) = 0$ and the matrix $\nabla_y h(\bar{x}, \bar{y})$ is nonsingular. Then there exist scalars $\varepsilon > 0$ and $\delta > 0$ and a function $\phi: S(\bar{x}; \varepsilon) \to S(\bar{y}; \delta)$ such that $\phi \in C^p$ over $S(\bar{x}; \varepsilon)$, $\bar{y} = \phi(\bar{x})$, and $h[x, \phi(x)] = 0$ for all $x \in S(\bar{x}; \varepsilon)$. The function

ϕ is unique in the sense that if $x \in S(\bar{x}; \varepsilon)$, $y \in S(\bar{y}; \delta)$, and $h(x, y) = 0$, then $y = \phi(x)$. Furthermore, if $p \geq 1$, then for all $x \in S(\bar{x}; \varepsilon)$

$$\nabla\phi(x) = -\nabla_x h[x, \phi(x)][\nabla_y h[x, \phi(x)]]^{-1}.$$

We shall also need the following implicit function theorem. It is a special case of a more general theorem found in Hestenes (1966). The notation (3) is used in the statement of the theorem.

Implicit Function Theorem 2: Let S be an open subset of R^{m+n}, $\overline{X}$ be a compact subset of R^m, and $h: S \to R^n$ be a function such that for some $p \geq 0$, $h \in C^p$ on S. Assume that $\nabla_y h(x, y)$ exists and is continuous on S. Assume that $\bar{y} \in R^n$ is a vector such that $(\bar{x}, \bar{y}) \in S$, $h(\bar{x}, \bar{y}) = 0$, and the matrix $\nabla_y h(\bar{x}, \bar{y})$ is nonsingular for all $\bar{x} \in \overline{X}$. Then there exist scalars $\varepsilon > 0$, $\delta > 0$, and a function $\phi: S(\overline{X}; \varepsilon) \to S(\bar{y}; \delta)$ such that $\phi \in C^p$ on $S(\overline{X}; \varepsilon)$, $\bar{y} = \phi(\bar{x})$ for all $\bar{x} \in \overline{X}$, and $h[x, \phi(x)] = 0$ for all $x \in S(\overline{X}; \varepsilon)$. The function ϕ is unique in the sense that if $x \in S(\overline{X}; \varepsilon)$, $y \in S(\bar{y}; \delta)$, and $h(x, y) = 0$, then $y = \phi(x)$. Furthermore, if $p \geq 1$, then for all $x \in S(\overline{X}; \varepsilon)$

$$\nabla\phi(x) = -\nabla_x h[x, \phi(x)][\nabla_y h[x, \phi(x)]]^{-1}.$$

When $\overline{X}$ consists of a single vector $\bar{x}$, the two implicit function theorems coincide.

Convexity

A set $S \subset R^n$ is said to be *convex* if for every $x, y \in S$ and $\alpha \in [0, 1]$ we have $\alpha x + (1 - \alpha)y \in S$. A function $f: S \to R$ is said to be *convex over the convex set* S if for every $x, y \in S$ and $\alpha \in [0, 1]$ we have

$$f[\alpha x + (1 - \alpha)y] \leq \alpha f(x) + (1 - \alpha)f(y).$$

If f is convex and $f \in C^1$ over an open convex set S, then

$$f(y) \geq f(x) + \nabla f(x)'(y - x) \qquad \forall\, x, y \in S. \tag{4}$$

If in addition $f \in C^2$ over S, then $\nabla^2 f(x) \geq 0$ for all $x \in S$. Conversely, if $f \in C^1$ over S and (4) holds, or if $f \in C^2$ over S and $\nabla^2 f(x) \geq 0$ for all $x \in S$, then f is convex over S.

Rate of Convergence Concepts

In minimization algorithms we are often interested in the speed with which various algorithms converge to a limit. Given a sequence $\{x_k\} \subset R^n$ with $x_k \to x^*$, the typical approach is to measure speed of convergence in terms of an *error function* $e: R^n \to R$ satisfying $e(x) \geq 0$ for all $x \in R^n$ and $e(x^*) = 0$. Typical choices are

$$e(x) = |x - x^*|, \qquad e(x) = |f(x) - f(x^*)|,$$

where f is the objective function of the problem. The sequence $\{e(x_k)\}$ is then compared with standard sequences. In our case, we compare $\{e(x_k)\}$ with geometric progressions of the form

$$r_k = q\beta^k,$$

where $q > 0$ and $\beta \in (0, 1)$ are some scalars, and with sequences of the form

$$r_k = q\beta^{p^k},$$

where $q > 0$, $\beta \in (0, 1)$, and $p > 1$ are some scalars. There is no reason for selecting these particular sequences for comparison other than the fact that they represent a sufficiently wide class which is adequate and convenient for our purposes. Our approach has much in common with that of Ortega and Rheinboldt (1970), except that we do not emphasize the distinction between Q and R linear or superlinear convergence.

Let us introduce some terminology:

Definition: Given two scalar sequences $\{e_k\}$ and $\{r_k\}$ with

$$0 \leq e_k, \qquad 0 \leq r_k, \qquad e_k \to 0, \qquad r_k \to 0,$$

we say that $\{e_k\}$ *converges faster than* $\{r_k\}$ if there exists an index $\bar{k} \geq 0$ such that

$$0 \leq e_k \leq r_k \qquad \forall\, k \geq \bar{k}.$$

We say that $\{e_k\}$ *converges slower than* $\{r_k\}$ if there exists an index $\bar{k} \geq 0$ such that

$$0 \leq r_k \leq e_k \qquad \forall\, k \geq \bar{k}.$$

Definition: Consider a scalar sequence $\{e_k\}$ with $e_k \geq 0$, $e_k \to 0$. The sequence $\{e_k\}$ is said to converge *at least linearly with convergence ratio* β, where $0 < \beta < 1$, if it converges faster than all geometric progressions of the form $q\bar{\beta}^k$ where $q > 0$, $\bar{\beta} \in (\beta, 1)$. It is said to converge *at most linearly with convergence ratio* β, where $0 < \beta < 1$, if it converges slower than all geometric progressions of the form $q\bar{\beta}^k$, where $q > 0$, $\bar{\beta} \in (0, \beta)$. It is said to converge *linearly with convergence ratio* β, where $0 < \beta < 1$, if it converges both at least and at most linearly with convergence ratio β. It is said to converge *superlinearly* or *sublinearly* if it converges faster or slower, respectively, then every sequence of the form $q\beta^k$, where $q > 0$, $\beta \in (0, 1)$.

Examples: (1) The following sequences all converge linearly with convergence ratio β:

$$q\beta^k, \quad q\left(\beta + \frac{1}{k}\right)^k, \quad q\left(\beta - \frac{1}{k}\right)^k, \quad q\beta^{k+(1/k)},$$

where $q > 0$ and $\beta \in (0, 1)$. This fact follows either by straightforward verification of the definition or by making use of Proposition 1.1 below.

(2) Let $0 < \beta_1 < \beta_2 < 1$, and consider the sequence $\{e_k\}$ defined by

$$e_{2k} = \beta_1^k \beta_2^k, \qquad e_{2k+1} = \beta_1^{k+1} \beta_2^k.$$

Then clearly $\{e_k\}$ converges at least linearly with convergence ratio β_2 and at most linearly with convergence ratio β_1. Actually $\{e_k\}$ can be shown to converge linearly with convergence ratio $\sqrt{\beta_1 \beta_2}$ a fact that can be proved by making use of the next proposition.

(3) The sequence $\{1/k\}$ converges sublinearly and every sequence of the form $q\beta^{p^k}$, where $q > 0$, $\beta \in (0, 1)$, $p > 1$, can be shown to converge superlinearly. Again these facts follow by making use of the proposition below.

Proposition 1.1: Let $\{e_k\}$ be a scalar sequence with $e_k \geq 0$, $e_k \to 0$. Then the following hold true:

(a) The sequence $\{e_k\}$ converges at least linearly with convergence ratio $\beta \in (0, 1)$ if and only if

$$\limsup_{k \to \infty} e_k^{1/k} \leq \beta. \tag{5}$$

It converges at most linearly with convergence ratio $\beta \in (0, 1)$ if and only if

$$\liminf_{k \to \infty} e_k^{1/k} \geq \beta. \tag{6}$$

It converges linearly with convergence ratio $\beta \in (0, 1)$ if and only if

$$\lim_{k \to \infty} e_k^{1/k} = \beta. \tag{7}$$

(b) If $\{e_k\}$ converges faster (slower) than some geometric progression of the form $q\beta^k$, $q > 0$, $\beta \in (0, 1)$, then it converges at least (at most) linearly with convergence ratio β.

(c) Assume that $e_k \neq 0$ for all k, and denote

$$\beta_1 = \liminf_{k \to \infty} \frac{e_{k+1}}{e_k}, \qquad \beta_2 = \limsup_{k \to \infty} \frac{e_{k+1}}{e_k}.$$

If $0 < \beta_1 < \beta_2 < 1$, then $\{e_k\}$ converges at least linearly with convergence ratio β_1 and at most linearly with convergence ratio β_2.

(d) Assume that $e_k \neq 0$ for all k and that

$$\lim_{k \to \infty} \frac{e_{k+1}}{e_k} = \beta.$$

If $0 < \beta < 1$, then $\{e_k\}$ converges linearly with convergence ratio β. If $\beta = 0$, then $\{e_k\}$ converges superlinearly. If $\beta = 1$, then $\{e_k\}$ converges sublinearly.

Proof: (a) If (5) holds, then for every $\bar{\beta} \in (\beta, 1)$ there exists a $\bar{k} \geq 0$ such that $e_k \leq \bar{\beta}^k$ for all $k \geq \bar{k}$. Since $\{\bar{\beta}^k\}$ converges faster than every sequence of the form $q\tilde{\beta}^k$, with $q > 0$, $\tilde{\beta} \in (\bar{\beta}, 1)$, the same is true for $\{e_k\}$. Since $\bar{\beta}$ can be taken arbitrarily close to β, it follows that $\{e_k\}$ converges at least linearly with convergence ratio β. Conversely if $\{e_k\}$ converges at least linearly with convergence ratio β, we have for every $\bar{\beta} \in (\beta, 1)$, $e_k \leq \bar{\beta}^k$ for all k sufficiently large. Hence, $\limsup_{k\to\infty} e_k^{1/k} \leq \bar{\beta}$. Since $\bar{\beta}$ can be taken arbitrarily close to β, (5) follows. An entirely similar argument proves the statement concerning (6). The statement regarding (7) is obtained by combining the two statements concerning (5) and (6).

(b) If $e_k \leq (\geq) q\beta^k$ for all k sufficiently large then $e_k^{1/k} \leq (\geq) q^{1/k}\beta$ and $\limsup_{k\to\infty} (\liminf_{k\to\infty}) e_k^{1/k} \leq (\geq)\beta$. Hence, by part (a), $\{e_k\}$ converges at least (at most) linearly with convergence ratio β.

(c) For every $\bar{\beta}_2 \in (\beta_2, 1)$, there exists $\bar{k} \geq 0$ such that

$$e_{k+1}/e_k \leq \bar{\beta}_2 \qquad \forall\, k \geq \bar{k}.$$

Hence, $e_{\bar{k}+m} \leq \bar{\beta}_2^m e_{\bar{k}}$ and $e_{\bar{k}+m}^{1/(\bar{k}+m)} \leq \bar{\beta}_2^{m/(\bar{k}+m)} e_{\bar{k}}^{1/(\bar{k}+m)}$. Taking the limit superior as $m \to \infty$, we obtain

$$\limsup_{k\to\infty} e_k^{1/k} \leq \bar{\beta}_2.$$

Since $\bar{\beta}_2$ can be taken arbitrarily close to β_2 we obtain $\limsup_{k\to\infty} e_k^{1/k} \leq \beta_2$, and the result follows by part (a). Similarly we prove the result relating to β_1.

(d) If $0 < \beta < 1$, the result follows directly from part (c). If $\beta = 0$, then for any $\bar{\beta} \in (0, 1)$ we have, for some $\bar{k} \geq 0$, $e_{k+1} \leq \bar{\beta} e_k$ for all $k \geq \bar{k}$. From this, it follows that $\{e_k\}$ converges faster than $\{\bar{\beta}^k\}$, and since $\bar{\beta}$ can be taken arbitrarily close to zero, $\{e_k\}$ converges superlinearly. Similarly we prove the result concerning sublinear convergence. Q.E.D.

When $\{e_k\}$ satisfies $\limsup_{k\to\infty} e_{k+1}/e_k = \beta < 1$ as in Proposition 1.1d, we also say that $\{e_k\}$ converges *at least quotient-linearly* (*or Q-linearly*) *with convergence ratio* β. If $\beta = 0$, then we say that $\{e_k\}$ converges *Q-superlinearly*.

Most optimization algorithms which are of interest in practice produce sequences converging either linearly or superlinearly. Linear convergence is quite satisfactory for optimization algorithms provided the convergence ratio β is not very close to unity. Algorithms which may produce sequences having sublinear convergence rates are excluded from consideration in most optimization problems as computationally inefficient. Several optimization algorithms possess superlinear convergence for particular classes of problems. For this reason, it is necessary to quantify further the notion of superlinear convergence.

Definition: Consider a scalar sequence $\{e_k\}$ with $e_k \geq 0$ converging superlinearly to zero. Then $\{e_k\}$ is said to *converge at least superlinearly with order p*, where $1 < p$, if it converges faster than all sequences of the form $q\beta^{\bar{p}^k}$, where $q > 0$, $\beta \in (0, 1)$, and $\bar{p} \in (1, p)$. It is said to converge *at most superlinearly with order p*, where $1 < p$, if it converges slower than all sequences of the form $q\beta^{\bar{p}^k}$, where $q > 0$, $\beta \in (0, 1)$, and $\bar{p} > p$. It is said to *converge superlinearly with order p*, where $p > 1$, if it converges both at least and at most superlinearly with order p.

We have the following proposition, the proof of which is similar to the one of Proposition 1.1 and is left as an exercise to the reader.

Proposition 1.2: Let $\{e_k\}$ be a scalar sequence with $e_k \geq 0$ and $e_k \to 0$. Then the following hold true:

(a) The sequence $\{e_k\}$ converges at least superlinearly with order $p > 1$ if and only if

$$\lim_{k \to \infty} e_k^{1/\bar{p}^k} = 0 \qquad \forall\, \bar{p} \in (1, p).$$

It converges at most superlinearly with order $p > 1$ if and only if

$$\lim_{k \to \infty} e_k^{1/\bar{p}^k} = 1 \qquad \forall\, \bar{p} > p.$$

(b) If $\{e_k\}$ converges faster (slower) than some sequence of the form $q\beta^{p^k}$, where $q > 0$, $\beta \in (0, 1)$, and $p > 1$, then it converges at least (at most) superlinearly with order p.

(c) Assume that $e_k \neq 0$ for all k. If for some $p > 1$, we have

$$\limsup_{k \to \infty} \frac{e_{k+1}}{e_k^p} < \infty,$$

then $\{e_k\}$ converges at least superlinearly with order p. If

$$\liminf_{k \to \infty} \frac{e_{k+1}}{e_k^p} > 0,$$

then $\{e_k\}$ converges at most superlinearly with order p.

If

$$\limsup_{k \to \infty} \frac{e_{k+1}}{e_k^p} < \infty,$$

as in Proposition 1.2c, then we say that $\{e_k\}$ converges *at least Q-superlinearly with order p*.

Cholesky Factorization

Let $A = [a_{ij}]$ be an $n \times n$ positive definite matrix and let us denote by A_i the *leading principal submatrix of A of order i*, $i = 1, \ldots, n$, where

$$A_i = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1i} \\ a_{21} & a_{22} & \cdots & a_{2i} \\ \vdots & \vdots & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ii} \end{bmatrix}.$$

It is easy to show that each of the submatrices A_i is a positive definite matrix. Indeed for any $y \in R^i$, $y \neq 0$, we have by positive definiteness of A

$$y'A_i y = [y' \quad 0]A\begin{bmatrix} y \\ 0 \end{bmatrix} > 0,$$

which implies that A_i is positive definite.

The matrices A_i satisfy

$$A_1 = [a_{11}],$$

$$A_i = \begin{bmatrix} A_{i-1} & \alpha_i \\ \alpha_i' & a_{ii} \end{bmatrix}, \qquad i = 2, \ldots, n, \tag{8}$$

where α_i is the column vector in R^{i-1} given by

$$\alpha_i = \begin{bmatrix} a_{1i} \\ \vdots \\ a_{i-1,i} \end{bmatrix}. \tag{9}$$

We now show that A can be written as

$$A = LL',$$

where L is a unique lower triangular matrix and L' is the transpose of L—an upper triangular matrix. This factorization of A is called the *Cholesky factorization.*

The Cholesky factorization may be obtained by successively factoring the principal submatrices A_i as

$$A_i = L_i L_i', \qquad i = 1, 2, \ldots, n. \tag{10}$$

We have

$$A_1 = L_1 L_1', \qquad L_1 = [\sqrt{a_{11}}].$$

Direct calculation using (8) yields that if $A_{i-1} = L_{i-1}L_{i-1}'$, then we also have $A_i = L_iL_i'$, where

$$L_i = \begin{bmatrix} L_{i-1} & 0 \\ l_i' & \lambda_{ii} \end{bmatrix}, \tag{11}$$

$$l_i = L_{i-1}^{-1}\alpha_i, \tag{12}$$

$$\lambda_{ii} = \sqrt{a_{ii} - l_i'l_i}, \tag{13}$$

and α_i is given by (9). Thus, to show that the factorization given above is valid, it will be sufficient to show that

$$a_{ii} - l_i'l_i > 0,$$

and thus λ_{ii} is well defined as a real number from (13). Indeed define $b = A_{i-1}^{-1}\alpha_i$. Then because A_i is positive definite, we have

$$\begin{aligned} 0 < [b' \quad -1]A_i\begin{bmatrix} b \\ -1 \end{bmatrix} &= b'A_{i-1}b - 2b'\alpha_i + a_{ii} \\ &= b'\alpha_i - 2b'\alpha_i + a_{ii} = a_{ii} - b'\alpha_i \\ &= a_{ii} - \alpha_i'A_{i-1}^{-1}\alpha_i = a_{ii} - \alpha_i'(L_{i-1}L_{i-1}')^{-1}\alpha_i \\ &= a_{ii} - (L_{i-1}^{-1}\alpha_i)'(L_{i-1}^{-1}\alpha_i) = a_{ii} - l_i'l_i. \end{aligned}$$

Thus, λ_{ii} as defined by (13) is well defined as a positive real number. In order to show uniqueness of the factorization, a similar induction argument may be used. The matrix A_1 has a unique factorization, and if A_{i-1} has a unique factorization $A_{i-1} = L_{i-1}L_{i-1}'$, then L_i is uniquely determined by the requirement $A_i = L_iL_i'$ and Eqs. (8)–(13).

In practice the Cholesky factorization is computed via the algorithm (10)–(13) or some other essentially equivalent algorithm. Naturally the vectors l_i in (12) are computed by solving the triangular system

$$L_{i-1}l_i = \alpha_i$$

rather than by inverting the matrix L_{i-1}. For large n the process requires approximately $n^3/6$ multiplications.

1.3 Unconstrained Minimization

We provide an overview of analytical and computational methods for solution of the problem

(UP) minimize $f(x)$

subject to $x \in R^n$,

where f: $R^n \to R$ is a given function. We say that a vector x^* is a *local minimum* for (UP) if there exists an $\varepsilon > 0$ such that

$$f(x^*) \leq f(x) \qquad \forall\, x \in S(x^*; \varepsilon).$$

It is a *strict local minimum* if there exists an $\varepsilon > 0$ such that

$$f(x^*) < f(x) \qquad \forall\, x \in S(x^*; \varepsilon), \quad x \neq x^*.$$

We have the following well-known optimality conditions. Proofs may be found, for example, in Luenberger (1973).

Proposition 1.3: Assume that x^* is a local minimum for (UP) and, for some $\varepsilon > 0, f \in C^1$ over $S(x^*; \varepsilon)$. Then

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

If in addition $f \in C^2$ over $S(x^*; \varepsilon)$, then

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

In what follows, we refer to a vector x^* satisfying $\nabla f(x^*) = 0$ as a *critical point.*

Proposition 1.4: Let x^* be such that, for some $\varepsilon > 0, f \in C^2$ over $S(x^*; \varepsilon)$ and

$$\nabla f(x^*) = 0, \qquad \nabla^2 f(x^*) > 0.$$

Then x^* is a strict local minimum for (UP). In fact, there exist scalars $\gamma > 0$ and $\delta > 0$ such that

$$f(x) \geq f(x^*) + \gamma |x - x^*|^2 \qquad \forall\, x \in S(x^*; \delta).$$

When x^* satisfies the assumptions of Proposition 1.4 we say that it is a *strong local minimum* for (UP).

We say that x^* is a *global minimum* for (UP) if

$$f(x^*) \leq f(x) \qquad \forall\, x \in R^n.$$

Under convexity assumptions on f, we have the following necessary and sufficient condition:

Proposition 1.5: Assume that $f \in C^1$ and is convex over R^n. Then a vector x^* is a global minimum for (UP) if and only if

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

Existence of global minima can be guaranteed under the assumptions of the following proposition which is a direct consequence of Weierstrass' theorem (a continuous function attains a global minimum over a compact set).

Proposition 1.6: If f is continuous over R^n and $f(x_k) \to \infty$ for every sequence $\{x_k\}$ such that $|x_k| \to \infty$, or, more generally, if the set $\{x \mid f(x) \leq \alpha\}$ is nonempty and compact for some $\alpha \in R$, then there exists a global minimum for (UP).

1.3.1 Convergence Analysis of Gradient Methods

We assume, without further mention throughout the remainder of Section 1.3, *that* $f \in C^1$ *over* R^n. The reader can easily make appropriate adjustments if $f \in C^1$ over an open subset of R^n only.

Most of the known iterative algorithms for solving (UP) take the form

$$x_{k+1} = x_k + \alpha_k d_k,$$

where if $\nabla f(x_k) \neq 0$, d_k is a *descent direction*, i.e., satisfies

$$d_k' \nabla f(x_k) < 0 \qquad \text{if} \quad \nabla f(x_k) \neq 0,$$

$$d_k = 0 \qquad \text{if} \quad \nabla f(x_k) = 0.$$

The scalar α_k is a positive stepsize parameter. We refer to such an algorithm as a *generalized gradient method* (or simply *gradient method*). Specific gradient methods that we shall consider include the method of steepest descent $[d_k = -\nabla f(x_k)]$ and scaled versions of it, Newton's method, the conjugate gradient method, quasi-Newton methods, and variations thereof. We shall examine several such methods in this section. For the time being, we focus on the convergence behavior of gradient methods. Rate of convergence issues will be addressed in the next subsection.

Stepsize Selection and Global Convergence

There are a number of rules for choosing the stepsize α_k [assuming $\nabla f(x_k) \neq 0$]. We list some that are used widely in practice:

(a) *Minimization rule:* Here α_k is chosen so that

$$f(x_k + \alpha_k d_k) = \min_{\alpha \geq 0} f(x_k + \alpha d_k).$$

(b) *Limited minimization rule:* A fixed number $s > 0$ is selected and α_k is chosen so that

$$f(x_k + \alpha_k d_k) = \min_{\alpha \in [0, s]} f(x_k + \alpha d_k).$$

(c) *Armijo rule:* Fixed scalars s, β, and σ with $s > 0$, $\beta \in (0, 1)$, and $\sigma \in (0, \frac{1}{2})$ are selected, and we set $\alpha_k = \beta^{m_k} s$, where m_k is the first nonnegative integer m for which

$$f(x_k) - f(x_k + \beta^m s d_k) \geq -\sigma \beta^m s \nabla f(x_k)' d_k,$$

i.e., $m = 0, 1, \ldots$ are tried successively until the inequality above is satisfied for $m = m_k$. (A variation of this rule is to use, instead of a fixed initial stepsize s, a sequence $\{s_k\}$ with $s_k > 0$ for all k. But this case can be reduced to the case of a fixed stepsize s by redefining the direction d_k to be $\tilde{d}_k = (s_k/s)d_k$.)

(d) *Goldstein rule:* A fixed scalar $\sigma \in (0, \frac{1}{2})$ is selected, and α_k is chosen to satisfy

$$\sigma \leq \frac{f(x_k + \alpha_k d_k) - f(x_k)}{\alpha_k \nabla f(x_k)' d_k} \leq 1 - \sigma.$$

It is possible to show that if f is bounded below there exists an interval of stepsizes α_k for which the relation above is satisfied, and there are fairly simple algorithms for finding such a stepsize through a finite number of arithmetic operations. However the Goldstein rule is primarily used in practice in conjunction with minimization rules in a scheme whereby an initial trial stepsize is chosen and tested to determine whether it satisfies the relation above. If it does, it is accepted. If not, a (perhaps approximate) line minimization is performed.

(e) *Constant stepsize:* Here a fixed stepsize $s > 0$ is selected and

$$\alpha_k = s \qquad \forall\, k.$$

The minimization and limited minimization rules must be implemented with the aid of one-dimensional line search algorithms (see, e.g., Luenberger, 1973; Avriel, 1976). In general, one cannot compute exactly the minimizing stepsize, and in practice, the line search is stopped once a stepsize α_k satisfying some termination criterion is obtained. An example of such a criterion is that α_k satisfies simultaneously

$$f(x_k) - f(x_k + \alpha_k d_k) \geq -\sigma \alpha_k \nabla f(x_k)' d_k \tag{1}$$

and

$$|\nabla f(x_k + \alpha_k d_k)' d_k| \leq \beta |\nabla f(x_k)' d_k|, \tag{2}$$

where σ and β are some scalars with $\sigma \in (0, \frac{1}{2})$ and $\beta \in (\sigma, 1)$. If α_k is indeed a minimizing stepsize then $\nabla f(x_k + \alpha_k d_k)' d_k = \partial f(x_k + \alpha_k d_k)/\partial \alpha = 0$, so (2) is in effect a test on the accuracy of the minimization. Relation (1), in view of $\nabla f(x_k)' d_k < 0$, guarantees a function decrease. Usually σ is chosen very close to zero, for example $\sigma \in [10^{-5}, 10^{-1}]$, but trial and error must be relied upon for the choice of β. Sometimes (2) is replaced by the less stringent condition

$$\nabla f(x_k + \alpha_k d_k)' d_k \geq \beta \nabla f(x_k)' d_k. \tag{3}$$

The following lemma shows that under mild assumptions there is an interval of stepsizes α satisfying (1), (2) or (1), (3).

Lemma 1.7: Assume that there is a scalar M such that $f(x) \geq M$ for all $x \in R^n$, let $\sigma \in (0, \frac{1}{2})$ and $\beta \in (\sigma, 1)$, and assume that $\nabla f(x_k)'d_k < 0$. There exists an interval $[c_1, c_2]$ with $0 < c_1 < c_2$, such that every $\alpha \in [c_1, c_2]$ satisfies (1) and (2) [and hence also (1) and (3)].

Proof: Define $g(\alpha) = f(x_k + \alpha d_k)$. Note that $\partial g(\alpha)/\partial \alpha = \nabla f(x_k + \alpha d_k)'d_k$. Let $\hat{\beta}$ be such that $\sigma < \hat{\beta} < \beta$, and consider the set A defined by

$$A = \left\{ \alpha \geq 0 \,\middle|\, \hat{\beta} \frac{\partial g(0)}{\partial \alpha} \leq \frac{\partial g(\alpha)}{\partial \alpha} \leq 0 \right\}.$$

Since $g(\alpha)$ is bounded below and $\partial g(0)/\partial \alpha = \nabla f(x_k)'d_k < 0$ it is easily seen that A is nonempty. Let

$$\hat{\alpha} = \min\{\alpha \mid \alpha \in A\}.$$

Clearly $\hat{\alpha} > 0$ and it is easy to see using the fact $\hat{\beta} < \beta$ that

$$\frac{\partial g(\alpha)}{\partial \alpha} \leq \hat{\beta} \frac{\partial g(0)}{\partial \alpha} \leq 0, \qquad \forall\, \alpha \in [0, \hat{\alpha}], \tag{4}$$

and there exists a scalar $\delta_1 \in (0, \hat{\alpha})$ such that

$$\left| \frac{\partial g(\alpha)}{\partial \alpha} \right| = |\nabla f(x_k + \alpha d_k)'d_k| \leq \beta |\nabla f(x_k)'d_k|$$

$$= \beta \left| \frac{\partial g(0)}{\partial \alpha} \right|, \qquad \forall\, \alpha \in [\hat{\alpha} - \delta_1, \hat{\alpha} + \delta_1].$$

We have from (4)

$$g(\hat{\alpha}) = g(0) + \int_0^{\hat{\alpha}} \frac{\partial g(t)}{\partial \alpha} dt \leq g(0) + \hat{\beta}\hat{\alpha} \frac{\partial g(0)}{\partial \alpha} < g(0) + \sigma\hat{\alpha} \frac{\partial g(0)}{\partial \alpha},$$

or equivalently

$$f(x_k) - f(x_k + \hat{\alpha}d_k) > -\sigma\hat{\alpha}\nabla f(x_k)'d_k.$$

Hence there exists a scalar $\delta_2 \in (0, \hat{\alpha})$ such that

$$f(x_k) - f(x_k + \alpha d_k) \geq -\sigma\alpha\nabla f(x_k)'d_k, \qquad \forall\, \alpha \in [\hat{\alpha} - \delta_2, \hat{\alpha} + \delta_2].$$

Take $\delta = \min\{\delta_1, \delta_2\}$. Then for all α in the interval $[\hat{\alpha} - \delta, \hat{\alpha} + \delta]$ both inequalities (1) and (2) are satisfied. Q.E.D.

In practice a line search procedure may have to be equipped with various mechanisms that guarantee that a stepsize satisfying the termination criteria will indeed be obtained. We refer the reader to more specific literature for details. In all cases, it is important to have a reasonably good initial stepsize

(or equivalently to scale the direction d_k in a reasonable manner). We discuss this in the next paragraph within the context of the Armijo rule.

The Armijo rule is very easy to implement and requires only one gradient evaluation per iteration. The process by which α_k is determined is shown in Fig. 1.1. We start with the trial point $(x_k + sd_k)$ and continue with $(x_k + \beta sd_k)$, $(x_k + \beta^2 sd_k), \ldots$ until the first time that $\beta^m s$ falls within the set of stepsizes α satisfying the desired inequality. While this set need not be an interval, it will always contain an interval of the form $[0, \delta]$ with $\delta > 0$, provided $\nabla f(x_k)'d_k < 0$. For this reason the stepsize α_k chosen by the Armijo rule is well defined and will be found after a finite number of trial evaluations of the value of f at the points $(x_k + sd_k)$, $(x_k + \beta sd_k), \ldots$. Usually σ is chosen close to zero, for example, $\sigma \in [10^{-5}, 10^{-1}]$. The scalar β is usually chosen from $\frac{1}{2}$ to 10^{-1} depending on the confidence we have on the quality of the initial stepsize s. Actually one can always take $s = 1$ and multiply the direction d_k by a scaling factor. Many methods incorporate automatic scaling of the direction d_k, which makes $s = 1$ a good stepsize choice (compare with Proposition 1.15 and the discussion on rate of convergence later in this section). If a suitable scaling factor for d_k is not known, one may use various ad hoc schemes to determine one. A simple possibility is to select a point $\bar{\alpha}$ on the line $\{x_k + \alpha d_k \mid \alpha > 0\}$, evaluate $f(x_k + \bar{\alpha}d_k)$, and perform a quadratic

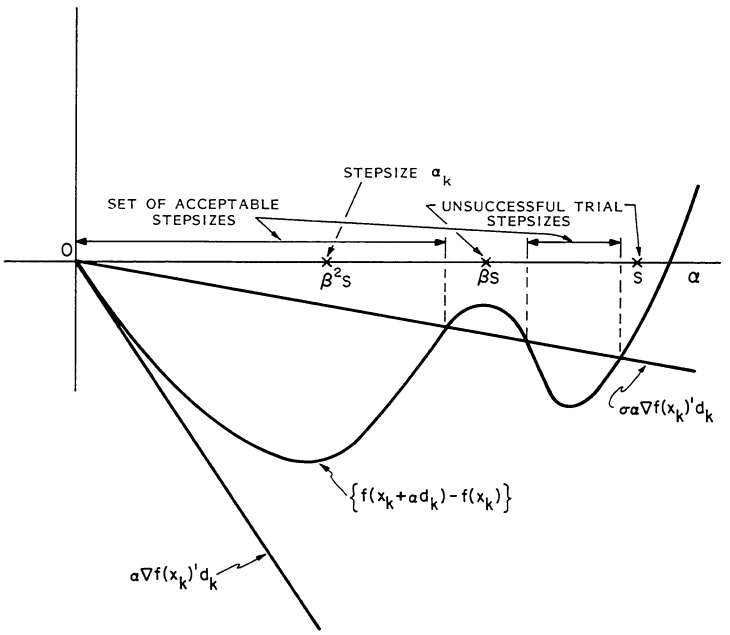

FIG. 1.1 Line search by the Armijo rule

interpolation on the basis of $f(x_k)$, $\nabla f(x_k)'d_k = \partial f(x_k + \alpha d_k)/\partial\alpha|_{\alpha=0}$, and $f(x_k + \bar{\alpha}d_k)$. If $\tilde{\alpha}$ minimizes the quadratic interpolation, d_k is replaced by $\tilde{d}_k = \tilde{\alpha}d_k$, and an initial stepsize $s = 1$ is used.

The constant stepsize rule is the simplest. It is useful in problems where evaluation of the objective function is expensive and an appropriate constant stepsize value is known or can be determined fairly easily. Interestingly enough, this is the case in the method of multipliers as we shall explain in the next chapter.

We now introduce a condition on the directions d_k of a gradient method.

Definition: Let $\{x_k\}$ be a sequence generated by a gradient method $x_{k+1} = x_k + \alpha_k d_k$. We say that the sequence $\{d_k\}$ is *uniformly gradient related* to $\{x_k\}$ if for every convergent subsequence $\{x_k\}_K$ for which

$$\lim_{\substack{k\to\infty \\ k\in K}} \nabla f(x_k) \neq 0 \tag{5}$$

there holds

$$0 < \liminf_{\substack{k\to\infty \\ k\in K}} |\nabla f(x_k)'d_k|, \qquad \limsup_{\substack{k\to\infty \\ k\in K}} |d_k| < \infty. \tag{6}$$

In words, $\{d_k\}$ is uniformly gradient related if whenever a subsequence $\{\nabla f(x_k)\}_K$ tends to a nonzero vector, the corresponding subsequence of directions d_k is bounded and does not tend to be orthogonal to $\nabla f(x_k)$. Another way of putting it is that (5) and (6) require that d_k does not become "too small" or "too large" relative to $\nabla f(x_k)$ and the angle between d_k and $\nabla f(x_k)$ does not get "too close" to $\pi/2$. Two examples of simple conditions that, if satisfied for some scalars $c_1 > 0$, $c_2 > 0$, $p_1 \geq 0$, and $p_2 \geq 0$ and all k, guarantee that $\{d_k\}$ is uniformly gradient related are

(a) $|d_k| \leq c_2|\nabla f(x_k)|^{p_2}, \qquad c_1|\nabla f(x_k)|^{p_1} \leq -\nabla f(x_k)'d_k;$

(b) $d_k = -D_k\nabla f(x_k),$

with D_k a positive definite symmetric matrix satisfying

$$c_1|\nabla f(x_k)|^{p_1}|z|^2 \leq z'D_k z \leq c_2|\nabla f(x_k)|^{p_2}|z|^2 \qquad \forall\, z \in R^n.$$

For example, in the method of steepest descent where $D_k = I$, this condition is satisfied if we take $c_1 = c_2 = 1$, $p_1 = p_2 = 0$.

We have the following convergence result:

Proposition 1.8: Let $\{x_k\}$ be a sequence generated by a gradient method $x_{k+1} = x_k + \alpha_k d_k$ and assume that $\{d_k\}$ is uniformly gradient related and α_k is chosen by the minimization rule, or the limited minimization rule, or the Armijo rule. Then every limit point of $\{x_k\}$ is a critical point.

Proof: Consider first the Armijo rule. Assume the contrary, i.e., that $\bar{x}$ is a limit point with $\nabla f(\bar{x}) \neq 0$. Then since $\{f(x_k)\}$ is monotonically decreasing and f is continuous, it follows that $\{f(x_k)\}$ converges to $f(\bar{x})$. Hence,

$$[f(x_k) - f(x_{k+1})] \to 0.$$

By the definition of the Armijo rule, we have

$$f(x_k) - f(x_{k+1}) \geq -\sigma\alpha_k \nabla f(x_k)' d_k.$$

Hence, $\alpha_k \nabla f(x_k)' d_k \to 0$. Let $\{x_k\}_K$ be the subsequence converging to $\bar{x}$. Since $\{d_k\}$ is uniformly gradient related, we have

$$\liminf_{\substack{k \to \infty \\ k \in K}} |\nabla f(x_k)' d_k| > 0,$$

and hence,

$$\{\alpha_k\}_K \to 0.$$

Hence, by the definition of the Armijo rule, we must have for some index $\bar{k} \geq 0$

$$(7) \quad f(x_k) - f[x_k + (\alpha_k/\beta)d_k] < -\sigma(\alpha_k/\beta)\nabla f(x_k)' d_k \qquad \forall\, k \in K, \quad k \geq \bar{k};$$

i.e., the initial stepsize s will be reduced at least once for all $k \in K$, $k \geq \bar{k}$. Denote

$$p_k = d_k/|d_k|, \qquad \bar{\alpha}_k = \alpha_k |d_k|/\beta.$$

Since $\{d_k\}$ is uniformly gradient related, we have $\limsup_{k \to \infty, k \in K} |d_k| < \infty$, and it follows that

$$\{\bar{\alpha}_k\}_K \to 0.$$

Since $|p_k| = 1$ for all $k \in K$, there exists a subsequence $\{p_k\}_{\bar{K}}$ of $\{p_k\}_K$ such that $\{p_k\}_{\bar{K}} \to \bar{p}$ where $\bar{p}$ is some vector with $|\bar{p}| = 1$. From (7), we have

$$(8) \quad \frac{f(x_k) - f(x_k + \bar{\alpha}_k p_k)}{\bar{\alpha}_k} < -\sigma \nabla f(x_k)' p_k \qquad \forall\, k \in \bar{K}, \quad k \geq \bar{k}.$$

Taking limits in (8) we obtain

$$-\nabla f(\bar{x})' \bar{p} \leq -\sigma \nabla f(\bar{x})' \bar{p} \qquad \text{or} \qquad 0 \leq (1 - \sigma)\nabla f(\bar{x})' \bar{p}.$$

Since $\sigma < 1$, we obtain

$$(9) \qquad 0 \leq \nabla f(\bar{x})' \bar{p}.$$

On the other hand, we have

$$-\nabla f(x_k)' p_k = -\nabla f(x_k)' d_k / |d_k|.$$

By taking the limit as $k \in \bar{K}$, $k \to \infty$,

$$-\nabla f(\bar{x})'\bar{p} \geq \frac{\liminf |\nabla f(x_k)'d_k|}{\limsup |d_k|} > 0,$$

which contradicts (9). This proves the result for the Armijo rule.

Consider now the minimization rule, and let $\{x_k\}_K$ converge to $\bar{x}$ with $\nabla f(\bar{x}) \neq 0$. Again we have that $\{f(x_k)\}$ decreases monotonically to $f(\bar{x})$. Let $\tilde{x}_{k+1}$ be the point generated from x_k via the Armijo rule, and let $\tilde{\alpha}_k$ be the corresponding stepsize. We have

$$f(x_k) - f(x_{k+1}) \geq f(x_k) - f(\tilde{x}_{k+1}) \geq -\sigma\tilde{\alpha}_k \nabla f(x_k)'d_k.$$

By simply replacing α_k by $\tilde{\alpha}_k$ and repeating the arguments of the earlier proof, we obtain a contradiction. In fact the line of argument just used establishes that *any stepsize rule that gives a larger reduction in objective function value at each step than the Armijo rule inherits its convergence properties.* This proves also the proposition for the limited minimization rule. Q.E.D.

Similarly the following proposition can be shown to be true. Its proof is left to the reader.

Proposition 1.9: The conclusions of Proposition 1.8 hold if $\{d_k\}$ is uniformly gradient related and α_k is chosen by the Goldstein rule or satisfies (1) and (2) for all k.

The next proposition establishes, among other things, convergence for the case of a constant stepsize.

Proposition 1.10: Let $\{x_k\}$ be a sequence generated by a gradient method $x_{k+1} = x_k + \alpha_k d_k$, where $\{d_k\}$ is uniformly gradient related. Assume that for some constant $L > 0$, we have

$$|\nabla f(x) - \nabla f(y)| \leq L|x - y| \qquad \forall\, x, y \in R^n, \tag{10}$$

and that there exists a scalar ε such that for all k we have $d_k \neq 0$ and

$$0 < \varepsilon \leq \alpha_k \leq \frac{2 - \varepsilon}{L} \frac{|\nabla f(x_k)'d_k|}{|d_k|^2}. \tag{11}$$

Then every limit point of $\{x_k\}$ is a critical point of f.

NOTE: If $\{d_k\}$ is such that there exist $c_1, c_2 > 0$ such that for all k we have

$$-\nabla f(x_k)'d_k \geq c_1 |\nabla f(x_k)|^2, \qquad c_2 |\nabla f(x_k)|^2 \geq |d_k|^2, \tag{12}$$

then (11) is satisfied if for all k we have

$$0 < \varepsilon \leq \alpha_k \leq (2 - \varepsilon)c_1/Lc_2. \tag{13}$$

For steepest descent $[d_k = -\nabla f(x_k)]$ in particular, we can take $c_1 = c_2 = 1$, and the condition on the stepsize becomes

$$0 < \varepsilon \leq \alpha_k \leq (2 - \varepsilon)/L.$$

Proof: We have the following equality for $\alpha \geq 0$,

$$f(x_k + \alpha d_k) = f(x_k) + \alpha \nabla f(x_k)' d_k + \int_0^\alpha [\nabla f(x_k + t d_k) - \nabla f(x_k)]' d_k \, dt.$$

By using (10), we obtain

$$\begin{aligned} f(x_k + \alpha d_k) - f(x_k) &\leq \alpha \nabla f(x_k)' d_k + \int_0^\alpha |\nabla f(x_k + t d_k) - \nabla f(x_k)| \, |d_k| \, dt \\ &\leq \alpha \nabla f(x_k)' d_k + \int_0^\alpha t L |d_k|^2 \, dt \\ &= \alpha[-|\nabla f(x_k)' d_k| + \tfrac{1}{2}\alpha L |d_k|^2]. \end{aligned}$$

From (11), we have $\alpha_k \geq \varepsilon$ and $\frac{1}{2}\alpha_k L |d_k|^2 - |\nabla f(x_k)' d_k| \leq -\frac{1}{2}\varepsilon |\nabla f(x_k)' d_k|$. Using these relations in the inequality above, we obtain

$$f(x_k) - f(x_k + \alpha_k d_k) \geq \tfrac{1}{2}\varepsilon^2 |\nabla f(x_k)' d_k|.$$

Now if a subsequence $\{x_k\}_K$ converges to a noncritical point $\bar{x}$, the above relation implies that $|\nabla f(x_k)' d_k| \to 0$. But this contradicts the fact that $\{d_k\}$ is uniformly gradient related. Hence, every limit point of $\{x_k\}$ is critical. Q.E.D.

Note that when $d_k = -D_k \nabla f(x_k)$ with D_k positive definite symmetric, relation (12) holds with

$$c_1 = \bar{\gamma}, \qquad c_2 = \bar{\Gamma}^2,$$

if the eigenvalues of D_k lie in the interval $[\bar{\gamma}, \bar{\Gamma}]$ for all k. It is also possible to show that (10) is satisfied for some $L > 0$, if $f \in C^2$ and the Hessian $\nabla^2 f$ is bounded over R^n. Unfortunately, however, it is difficult in general to obtain an estimate of L and thus in most cases the interval of stepsizes in (11) or (13) which guarantees convergence is not known a priori. Thus, experimentation with the problem at hand is necessary in order to obtain a range of stepsize values which lead to convergence. We note, however, that in the method of multipliers, it is possible to obtain a satisfactory estimate of L as will be explained in Chapter 2.

Gradient Convergence

The convergence results given so far are concerned with limit points of the sequence $\{x_k\}$. It can also be easily seen that the corresponding sequence $\{f(x_k)\}$ will converge to some value whenever $\{x_k\}$ has at least one limit point and there holds $f(x_{k+1}) \leq f(x_k)$ for all k. Concerning the sequence $\{\nabla f(x_k)\}$, we have by continuity of ∇f that if a subsequence $\{x_k\}_K$ converges to some point $\bar{x}$ then $\{\nabla f(x_k)\}_K \to \nabla f(\bar{x})$. If $\bar{x}$ is critical, then $\{\nabla f(x_k)\}_K \to 0$. More generally, we have the following result:

Proposition 1.11: Let $\{x_k\}$ be a sequence generated by a gradient method $x_{k+1} = x_k + \alpha_k d_k$, which is convergent in the sense that every limit point of sequences that it generates is a critical point of f. Then if $\{x_k\}$ is a bounded sequence, we have $\nabla f(x_k) \to 0$.

Proof: Assume the contrary, i.e., that there exists a subsequence $\{x_k\}_K$ and an $\varepsilon > 0$ such that $|\nabla f(x_k)| \geq \varepsilon$ for all $k \in K$. Since $\{x_k\}_K$ is bounded, it has at least one limit point $\bar{x}$ and we must have $|\nabla f(\bar{x})| \geq \varepsilon$. But this contradicts our hypothesis which implies that $\bar{x}$ must be critical. Q.E.D.

The proposition above forms the basis for terminating the iterations of gradient methods. Thus, computation is stopped when a point $x_{\bar{k}}$ is obtained with

$$|\nabla f(x_{\bar{k}})| \leq \varepsilon, \tag{14}$$

where ε is a small positive scalar. The point $x_{\bar{k}}$ is considered for practical purposes to be a critical point. Sometimes one terminates computation when the norm of the direction d_k becomes too small; i.e.,

$$|d_{\bar{k}}| \leq \varepsilon. \tag{15}$$

If d_k satisfies

$$c_1 |\nabla f(x_k)|^{p_1} \leq |d_k| \leq c_2 |\nabla f(x_k)|^{p_2}$$

for some positive scalars c_1, c_2, p_1, p_2, and all k, then the termination criterion (15) is of the same nature as (14). Unfortunately, it is not known a priori how small one should take ε in order to guarantee that the final point $x_{\bar{k}}$ is a "good" approximation to a stationary point. For this reason it is necessary to conduct some experimentation prior to settling on a reasonable termination criterion for a given problem, unless bounds are known (or can be estimated) for the Hessian matrix of f (see the following exercise).

Exercise: Let x^* be a local minimum of f and assume that for all x in a sphere $S(x^*; \delta)$ we have, for some $m > 0$ and $M > 0$,

$$m|z|^2 \leq z' \nabla^2 f(x) z \leq M|z|^2 \qquad \forall\, z \in R^n.$$

Then every $x \in S(x^*; \delta)$ satisfying $|\nabla f(x)| \leq \varepsilon$ also satisfies

$$|x - x^*| \leq \varepsilon/m, \qquad f(x) - f(x^*) \leq M\varepsilon^2/2m^2.$$

Local Convergence

A weakness of the convergence results of the preceding subsection is that they do not guarantee that convergence (to a single point) of the generated sequence $\{x_k\}$ will occur. Thus, the sequence $\{x_k\}$ may have one, more than one, or no limit points at all. It is not infrequent for a gradient method to generate an unbounded sequence $\{x_k\}$. This will typically occur if the function f has no critical point or if f decreases monotonically as $|x| \to \infty$ along some directions. However $\{x_k\}$ will have at least one limit point if the set $\{x \mid f(x) \leq f(x_0)\}$ is bounded or more generally if $\{x_k\}$ is a bounded sequence.

On the other hand, practical experience suggests that a sequence generated by a gradient method will rarely have more than one critical limit point. This is not very surprising since the generated sequence of function values $\{f(x_k)\}$ is monotonically nonincreasing and will always converge to a finite value whenever $\{x_k\}$ has at least one limit point. Hence, any two critical limit points, say $\bar{x}$ and $\tilde{x}$, of the sequence $\{x_k\}$ must simultaneously satisfy $\nabla f(\bar{x}) = \nabla f(\tilde{x}) = 0$ and $f(\bar{x}) = f(\tilde{x}) = \lim_{k\to\infty} f(x_k)$. These relations are unlikely to hold if the critical points of f are "isolated" points. One may also prove that if f has a *finite* number of critical points and the Armijo rule or the limited minimization rule is used in connection with a gradient method with uniformly gradient-related direction sequence $\{d_k\}$, then the generated sequence $\{x_k\}$ will converge to a unique critical point provided that $\{x_k\}$ is a bounded sequence. We leave this as an exercise for the reader.

The following proposition may also help to explain to some extent why sequences generated by gradient methods tend to have unique limit points. It states that strong local minima tend to attract gradient methods.

Proposition 1.12: Let $f \in C^2$ and $\{x_k\}$ be a sequence satisfying $f(x_{k+1}) \leq f(x_k)$ for all k and generated by a gradient method $x_{k+1} = x_k + \alpha_k d_k$ which is convergent in the sense that every limit point of sequences that it generates is a critical point of f. Assume that there exist scalars $s > 0$ and $c > 0$ such that for all k there holds $\alpha_k \leq s$ and $|d_k| \leq c|\nabla f(x_k)|$. Then for every local minimum x^* of f with $\nabla^2 f(x^*) > 0$, there exists an open set L containing x^* such that if $x_{\bar{k}} \in L$ for some $\bar{k} \geq 0$ then $x_k \in L$ for all $k \geq \bar{k}$ and $\{x_k\} \to x^*$. Furthermore, given any scalar $\varepsilon > 0$, the set L can be chosen so that $L \subset S(x^*; \varepsilon)$.

NOTE: The condition $\alpha_k \leq s$ is satisfied for the Armijo rule and the limited minimization rule. The condition $|d_k| \leq c|\nabla f(x_k)|$ is satisfied if $d_k = -D_k \nabla f(x_k)$ with the eigenvalues of D_k uniformly bounded from above.

Proof: Let x^* be a local minimum with $\nabla^2 f(x^*)$ positive definite. Then there exists $\bar{\varepsilon} > 0$ such that for all x with $|x - x^*| \leq \bar{\varepsilon}$, the matrix $\nabla^2 f(x)$ is also positive definite. Denote

$$\gamma = \min_{\substack{|x-x^*|\leq\bar{\varepsilon}\\|z|=1}} z'\nabla^2 f(x)z, \qquad \Gamma = \max_{\substack{|x-x^*|\leq\bar{\varepsilon}\\|z|=1}} z'\nabla^2 f(x)z.$$

We have $\gamma > 0$ and $\Gamma > 0$. Consider the open set

$$L = \{x \,|\, |x - x^*| < \bar{\varepsilon},\ f(x) < f(x^*) + \tfrac{1}{2}\gamma[\bar{\varepsilon}/(1 + sc\Gamma)]^2\}.$$

We claim that if $x_{\bar{k}} \in L$ for some $\bar{k} \geq 0$ then $x_k \in L$ for all $k \geq \bar{k}$ and furthermore $x_k \to x^*$.

Indeed if $x_{\bar{k}} \in L$ then by using Taylor's theorem, we have

$$\tfrac{1}{2}\gamma|x_{\bar{k}} - x^*|^2 \leq f(x_{\bar{k}}) - f(x^*) < \tfrac{1}{2}\gamma[\bar{\varepsilon}/(1 + sc\Gamma)]^2$$

from which we obtain

$$|x_{\bar{k}} - x^*| < \bar{\varepsilon}/(1 + cs\Gamma). \tag{16}$$

On the other hand, we have

$$\begin{aligned} |x_{\bar{k}+1} - x^*| &= |x_{\bar{k}} - x^* + \alpha_{\bar{k}} d_{\bar{k}}| \leq |x_{\bar{k}} - x^*| + \alpha_{\bar{k}}|d_{\bar{k}}| \\ &\leq |x_{\bar{k}} - x^*| + sc|\nabla f(x_{\bar{k}})|. \end{aligned}$$

By using Taylor's theorem, we have $|\nabla f(x_{\bar{k}})| \leq \Gamma|x_{\bar{k}} - x^*|$ and substituting in the inequality above, we obtain

$$|x_{\bar{k}+1} - x^*| \leq (1 + sc\Gamma)|x_{\bar{k}} - x^*|.$$

By combining this relation with (16), we obtain

$$|x_{\bar{k}+1} - x^*| < \bar{\varepsilon}.$$

Furthermore, using the hypothesis $f(x_{k+1}) \leq f(x_k)$ for all k, we have

$$f(x_{\bar{k}+1}) \leq f(x_{\bar{k}}) < f(x^*) + \tfrac{1}{2}\gamma[\bar{\varepsilon}/(1 + sc\Gamma)]^2.$$

It follows from the above two inequalities that $x_{\bar{k}+1} \in L$ and similarly $x_k \in L$ for all $k \geq \bar{k}$. Let $\bar{L}$ be the closure of L. Since $\bar{L}$ is a compact set, the sequence $\{x_k\}$ will have at least one limit point which by assumption must be a critical point of f. Now the only critical point of f within $\bar{L}$ is the point x^* (since f is strictly convex within $\bar{L}$). Hence $x_k \to x^*$. Finally given any $\varepsilon > 0$, we can choose $\bar{\varepsilon} \leq \varepsilon$ in which case $L \subset S(x^*; \varepsilon)$. Q.E.D.

Rate of Convergence—Quadratic Objective Function

The second major question relating to the behavior of a gradient method concerns the speed (or rate) of convergence of generated sequences $\{x_k\}$. The mere fact that x_k converges to a critical point x^* will be of little value in

practice unless the points x_k are reasonably close to x^* after relatively few iterations. Thus, the study of the rate of convergence of an algorithm or a class of algorithms not only provides useful information regarding computational efficiency, but also delineates what in most cases are the dominant criteria for selecting one algorithm in favor of others for solving a particular problem.

Most of the important characteristics of gradient methods are revealed by investigation of the case where the objective function is quadratic. Indeed, assume that a gradient method is applied to minimization of a function $f: R^n \to R$, $f \in C^2$, and it generates a sequence $\{x_k\}$ converging to a strong local minimum x^* where

$$\nabla f(x^*) = 0, \qquad \nabla^2 f(x^*) > 0.$$

Then we have, by Taylor's Theorem,

$$f(x) = f(x^*) + \tfrac{1}{2}(x - x^*)'\nabla^2 f(x^*)(x - x^*) + o(|x - x^*|^2),$$

where $o(|x - x^*|^2)/|x - x^*|^2 \to 0$ as $x \to x^*$. This implies that f can be accurately approximated near x^* by the quadratic function

$$f(x^*) + \tfrac{1}{2}(x - x^*)'\nabla^2 f(x^*)(x - x^*).$$

We thus expect that rate-of-convergence results obtained through analysis of the case where the objective function is the quadratic function above have direct analogs to the general case. The validity of this conjecture can indeed be established by rigorous analysis and has been substantiated by extensive numerical experimentation.

Consider the quadratic function

$$f(x) = \tfrac{1}{2}(x - x^*)'Q(x - x^*)$$

and the gradient method

$$x_{k+1} = x_k - \alpha_k D_k g_k, \tag{17}$$

where

$$g_k = \nabla f(x_k) = Q(x_k - x^*). \tag{18}$$

We assume that Q and D_k are positive definite and symmetric. Let

$$M_k = \text{max eigenvalue of } (D_k^{1/2} Q D_k^{1/2}),$$

$$m_k = \text{min eigenvalue of } (D_k^{1/2} Q D_k^{1/2}).$$

We have the following proposition:

Proposition 1.13: Consider iteration (17), and assume that α_k is chosen according to the minimization rule

$$f(x_k - \alpha_k D_k g_k) = \min_{\alpha \geq 0} f(x_k - \alpha D_k g_k).$$

Then

$$f(x_{k+1}) \leq \left(\frac{M_k - m_k}{M_k + m_k}\right)^2 f(x_k). \tag{19}$$

Proof: The result clearly holds if $g_k = 0$, so we assume $g_k \neq 0$. We first compute the minimizing stepsize α_k. We have

$$\begin{aligned}(d/d\alpha)f(x_k - \alpha D_k g_k) &= -g_k' D_k Q(x_k - \alpha D_k g_k - x^*) \\ &= -g_k' D_k g_k + \alpha g_k' D_k Q D_k g_k.\end{aligned}$$

Hence, by setting this derivative equal to zero, we obtain

$$\alpha_k = g_k' D_k g_k / g_k' D_k Q D_k g_k. \tag{20}$$

We have, using (17) and (20),

$$\begin{aligned}f(x_{k+1}) &= f(x_k - \alpha_k D_k g_k) = \tfrac{1}{2}(x_k - x^* - \alpha_k D_k g_k)' Q(x_k - x^* - \alpha_k D_k g_k) \\ &= \tfrac{1}{2}(x_k - x^*)' Q(x_k - x^*) + \tfrac{1}{2}\alpha_k^2 g_k' D_k Q D_k g_k - \alpha_k g_k' D_k Q(x_k - x^*) \\ &= f(x_k) + \tfrac{1}{2}\alpha_k^2 g_k' D_k Q D_k g_k - \alpha_k g_k' D_k g_k,\end{aligned}$$

and finally

$$f(x_{k+1}) = f(x_k) - \frac{1}{2}\frac{(g_k' D_k g_k)^2}{g_k' D_k Q D_k g_k}. \tag{21}$$

Also we have

$$\begin{aligned}f(x_k) &= \tfrac{1}{2}(x_k - x^*)' Q(x_k - x^*) \\ &= \tfrac{1}{2}(x_k - x^*)' Q D_k^{1/2}(D_k^{1/2} Q D_k^{1/2})^{-1} D_k^{1/2} Q(x_k - x^*) \\ &= \tfrac{1}{2} g_k' D_k^{1/2}(D_k^{1/2} Q D_k^{1/2})^{-1} D_k^{1/2} g_k.\end{aligned} \tag{22}$$

Setting $y_k = D_k^{1/2} g_k$, $L_k = D_k^{1/2} Q D_k^{1/2}$, and using (21) and (22), we obtain

$$\begin{aligned}f(x_{k+1}) &= f(x_k) - \frac{(y_k' y_k)^2}{(y_k' L_k y_k)(y_k' L_k^{-1} y_k)} f(x_k) \\ &= \left[1 - \frac{(y_k' y_k)^2}{(y_k' L_k y_k)(y_k' L_k^{-1} y_k)}\right] f(x_k).\end{aligned} \tag{23}$$

We shall now need the following lemma, a proof of which can be found in Luenberger (1973, p. 151).

Lemma (Kantorovich Inequality): Let L be a positive definite symmetric $n \times n$ matrix. Then for any vector $y \in R^n$, $y \neq 0$, there holds

$$\frac{(y'y)^2}{(y'Ly)(y'L^{-1}y)} \geq \frac{4Mm}{(M + m)^2},$$

where M and m are the largest and smallest eigenvalues of L.

Returning to the proof of the proposition, we have by using Kantorovich's inequality in (23)

$$f(x_{k+1}) \leq \left[1 - \frac{4M_k m_k}{(M_k + m_k)^2}\right] f(x_k) = \left(\frac{M_k - m_k}{M_k + m_k}\right)^2 f(x_k). \qquad \text{Q.E.D.}$$

From (19), we obtain, assuming $g_k \neq 0$ for all k,

$$\limsup_{k \to \infty} \frac{f(x_{k+1})}{f(x_k)} \leq \limsup_{k \to \infty} \left(\frac{M_k - m_k}{M_k + m_k}\right)^2 \triangleq \beta.$$

If $\beta < 1$ (as will be the case if $\{m_k/M_k\}$ is bounded away from zero), it follows that $\{f(x_k)\}$ converges at least Q-linearly with convergence ratio β (see Section 1.2). If $\beta = 0$, then the convergence rate is *superlinear*. If $\beta < 1$, then the sequence $\{f(x_{k+1})\}$ is majorized for all k sufficiently large by any geometric progression of the form $q\bar{\beta}^k$, where $q > 0$, $\bar{\beta} > \beta$ (see Section 1.2). If γ is the minimum eigenvalue of Q, we have

$$\tfrac{1}{2}\gamma |x_k - x^*|^2 \leq f(x_k)$$

so the same conclusion can be drawn for the sequence $\{|x_k - x^*|^2\}$. Relation (19) also indicates that the iteration $x_{k+1} = x_k - \alpha_k D_k g_k$ yields a large relative reduction in objective function value if $M_k/m_k \sim 1$. This shows that in order to achieve fast convergence, one should select D_k so that the eigenvalues of $D_k^{1/2} Q D_k^{1/2}$ are close together, such as when $D_k \sim Q^{-1}$, and this is the main motivation for introducing the matrix D_k instead of taking $D_k \equiv I$. If in particular $D_k = Q^{-1}$, then we obtain $M_k = m_k = 1$ and, from (19), $f(x_{k+1}) = 0$ which implies $x_{k+1} = x^*$; i.e., convergence to the minimum is attained in a single iteration.

When the ratio M_k/m_k is much larger than unity, then (19) indicates that convergence can be very slow. Actually, the speed of convergence of $\{x_k\}$ depends strongly on the starting point x_0. However, if D_k is constant, it is possible to show that there always exist "worst" starting points for which (19) is satisfied with equality for all k. [The reader may wish to verify this by considering the case $D_k \equiv I, f(x) = \frac{1}{2}\sum_{i=1}^n \gamma_i x_i^2$, where $0 < \gamma_1 \leq \gamma_2 \leq \cdots \leq \gamma_n$, and the starting point $x_0 = (\gamma_1^{-1}, 0, \ldots, 0, \gamma_n^{-1})$.]

Similar convergence rate results can be obtained for the case of the limited minimization rule. For example, notice that from (20), we obtain

$$\alpha_k = y_k' y_k / y' D_k^{1/2} Q D_k^{1/2} y_k,$$

where $y_k = D_k^{1/2} g_k$. Hence, we have $\alpha_k \leq 1/m_k$, and (19) also holds when α_k is chosen by the limited minimization rule

$$f(x_k - \alpha_k D_k g_k) = \min_{0 \leq \alpha \leq s} f(x_k - \alpha D_k g_k)$$

provided that

$$s \geq 1/m_k, \qquad k = 0, 1, \ldots .$$

Qualitatively, similar results are also obtained when other stepsize rules are used, such as a constant stepsize. We have the following proposition:

Proposition 1.14: Consider the iteration $x_{k+1} = x_k - \alpha_k D_k g_k$. For all $\alpha_k \geq 0$ and k, we have

$$(24) \qquad (x_{k+1} - x^*)' D_k^{-1}(x_{k+1} - x^*) \leq \max\{|1 - \alpha_k m_k|^2, |1 - \alpha_k M_k|^2\}(x_k - x^*)' D_k^{-1}(x_k - x^*).$$

Furthermore, the right-hand side of (24) is minimized when

$$(25) \qquad \alpha_k = 2/(m_k + M_k),$$

and with this choice of α_k, we obtain

$$(26) \qquad (x_{k+1} - x^*)' D_k^{-1}(x_{k+1} - x^*) \leq \left(\frac{M_k - m_k}{M_k + m_k}\right)^2 (x_k - x^*)' D_k^{-1}(x_k - x^*).$$

Proof: We have

$$x_{k+1} - x^* = x_k - x^* - \alpha_k D_k g_k = x_k - x^* - \alpha_k D_k Q(x_k - x^*).$$

A straightforward calculation yields

$$(x_{k+1} - x^*)' D_k^{-1}(x_{k+1} - x^*) = (x_k - x^*)' D_k^{-1/2}(I - \alpha_k D_k^{1/2} Q D_k^{1/2})^2 D_k^{-1/2}(x_k - x^*).$$

Hence,

$$(x_{k+1} - x^*)' D_k^{-1}(x_{k+1} - x^*) \leq A_k^2 (x_k - x^*)' D_k^{-1}(x_k - x^*),$$

where A_k is the maximum eigenvalue of $G_k = (I - \alpha_k D_k^{1/2} Q D_k^{1/2})$. The eigenvalues of G_k are $1 - \alpha_k e_i(D_k^{1/2} Q D_k^{1/2})$, $i = 1, \ldots, n$, where $e_i(D_k^{1/2} Q D_k^{1/2})$ is the ith eigenvalue of $D_k^{1/2} Q D_k^{1/2}$. From this we obtain by an elementary calculation

$$|A_k| = \max\{|1 - \alpha_k m_k|, |1 - \alpha_k M_k|\},$$

and (24) follows. The verification of the fact that α_k as given by (25) minimizes the right-hand side of (24) is elementary and is left to the reader. Q.E.D.

The result shows that if $D_k = D$ for all k where D is positive definite and

$$\limsup_{k \to \infty} \max\{|1 - \alpha_k m|^2, |1 - \alpha_k M|^2\} = \beta,$$

where m, M are the smallest and largest eigenvalues of $(D^{1/2} Q D^{1/2})$, then $\{(x_k - x^*)' D^{-1}(x_k - x^*)\}$ converges at least linearly with convergence

ratio β provided $0 < \beta < 1$. If $c > 0$ is the smallest eigenvalue of D^{-1} and Γ is the largest eigenvalue of Q, we have

$$(c/\Gamma)f(x_k) \leq \tfrac{1}{2}c|x_k - x^*|^2 \leq \tfrac{1}{2}(x_k - x^*)'D^{-1}(x_k - x^*).$$

Hence, if $0 < \beta < 1$, we have that $\{f(x_k)\}$ and $\{|x_k - x^*|^2\}$ will also converge faster than linearly with convergence ratio β. The important point is that [compare with (26)]

$$\left(\frac{M - m}{M + m}\right)^2 \leq \beta,$$

and hence if M/m is much larger than unity, again the convergence rate can be very slow even if the optimal stepsize $\alpha_k = 2/(m_k + M_k)$ (which is generally unknown) were to be utilized. From this, it follows again that D_k should be chosen as close as possible to Q^{-1} so that $M_k \sim m_k \sim 1$. Notice that if D_k has indeed been so chosen, then (25) shows that the stepsize $\alpha_k = 1$ is a good choice. This fact also follows from (20), which shows that when $D_k \sim Q^{-1}$ then the minimizing stepsize is near unity.

Rate of Convergence—Nonquadratic Objective Function

One can show that our main conclusions on rate of convergence carry over to the nonquadratic case for sequences converging to strong local minima.

Let $f \in C^2$ and consider the gradient method

$$x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k), \tag{27}$$

where D_k is positive definite symmetric. Consider a generated sequence $\{x_k\}$ and assume that

$$x_k \to x^*, \qquad \nabla f(x^*) = 0, \qquad \nabla^2 f(x^*) > 0, \tag{28}$$

and that $x_k \neq x^*$ for all k. Then it is possible to show the following:

(a) If α_k is chosen by the line minimization rule there holds

$$\limsup_{k\to\infty} \frac{f(x_{k+1}) - f(x^*)}{f(x_k) - f(x^*)} \leq \limsup_{k\to\infty} \left(\frac{M_k - m_k}{M_k + m_k}\right)^2, \tag{29}$$

where M_k and m_k are the largest and smallest eigenvalues of $D_k^{1/2}\nabla^2 f(x^*)D_k^{1/2}$.

(b) There holds

$$\limsup_{k\to\infty} \frac{(x_{k+1} - x^*)'D_k^{-1}(x_{k+1} - x^*)}{(x_k - x^*)'D_k^{-1}(x_k - x^*)} \leq \limsup_{k\to\infty} \max\{|1 - \alpha_k m_k|^2, |1 - \alpha_k M_k|^2\}.$$

The proof of these facts involves essentially a repetition of the proofs of Propositions 1.13 and 1.14. However, the details are somewhat more technical and will not be given.

When $D_k \to \nabla^2 f(x^*)^{-1}$, then (29) shows that the convergence rate of $\{f(x_k) - f(x^*)\}$ is superlinear. A somewhat more general version of this result for the case of the Armijo rule is given by the following proposition:

Proposition 1.15: Consider a sequence $\{x_k\}$ generated by (27) and satisfying (28). Assume further that $\nabla f(x_k) \neq 0$ for all k and

$$\lim_{k\to\infty} \frac{|[D_k - \nabla^2 f(x^*)^{-1}]\nabla f(x_k)|}{|\nabla f(x_k)|} = 0. \tag{30}$$

Then if α_k is chosen by means of the Armijo rule with initial stepsize $s = 1$, we have

$$\lim_{k\to\infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|} = 0,$$

and hence $\{|x_k - x^*|\}$ converges superlinearly. Furthermore, there exists an integer $\bar{k} \geq 0$ such that we have $\alpha_k = 1$ for all $k \geq \bar{k}$ (i.e., eventually no reduction of the initial stepsize will be taking place).

Proof: We first prove that there exists a $\bar{k} \geq 0$ such that for all $k \geq \bar{k}$ we have $\alpha_k = 1$. By the mean value theorem we have

$$f(x_k) - f[x_k - D_k\nabla f(x_k)] = \nabla f(x_k)'D_k\nabla f(x_k) - \tfrac{1}{2}\nabla f(x_k)'D_k\nabla^2 f(\bar{x}_k)D_k\nabla f(x_k),$$

where $\bar{x}_k$ is a point on the line segment joining x_k and $x_k - D_k\nabla f(x_k)$. It will be sufficient to show that for k sufficiently large we have

$$\nabla f(x_k)'D_k\nabla f(x_k) - \tfrac{1}{2}\nabla f(x_k)'D_k\nabla^2 f(\bar{x}_k)D_k\nabla f(x_k) \geq \sigma\nabla f(x_k)'D_k\nabla f(x_k)$$

or equivalently, by defining $p_k = \nabla f(x_k)/|\nabla f(x_k)|$,

$$(1 - \sigma)p_k'D_kp_k \geq \tfrac{1}{2}p_k'D_k\nabla^2 f(\bar{x}_k)D_kp_k. \tag{31}$$

From (28), (30), we obtain $D_k\nabla f(x_k) \to 0$. Hence, $x_k - D_k\nabla f(x_k) \to x^*$, and it follows that $\bar{x}_k \to x^*$ and $\nabla^2 f(\bar{x}_k) \to \nabla^2 f(x^*)$. Now (30) is written as

$$D_kp_k = [\nabla^2 f(x^*)]^{-1}p_k + \beta_k,$$

where $\{\beta_k\}$ denotes a vector sequence with $\beta_k \to 0$. By using the above relation and the fact that $\nabla^2 f(\bar{x}_k) \to \nabla^2 f(x^*)$, we may write (31) as

$$(1 - \sigma)p_k'[\nabla^2 f(x^*)]^{-1}p_k \geq \tfrac{1}{2}p_k'[\nabla^2 f(x^*)]^{-1}p_k + \gamma_k,$$

where $\{\gamma_k\}$ is some scalar sequence with $\gamma_k \to 0$. Thus (31) is equivalent to

$$(\tfrac{1}{2} - \sigma)p_k'[\nabla^2 f(x^*)]^{-1}p_k \geq \gamma_k.$$

Since $\frac{1}{2} - \sigma > 0$, $|p_k| = 1$, and $\nabla^2 f(x^*) > 0$, the above relation holds for k sufficiently large, and we have $\alpha_k = 1$ for $k \geq \bar{k}$ where $\bar{k}$ is some index.

To show superlinear convergence we write, for $k \geq \bar{k}$,

$$x_{k+1} - x^* = x_k - x^* - D_k \nabla f(x_k). \tag{32}$$

We have, from (30) and for some sequence $\{\delta_k\}$ with $\delta_k \to 0$,

$$D_k \nabla f(x_k) = \nabla^2 f(x^*)^{-1} \nabla f(x_k) + |\nabla f(x_k)| \delta_k. \tag{33}$$

From Taylor's theorem we obtain

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

from which

$$[\nabla^2 f(x^*)]^{-1} \nabla f(x_k) = x_k - x^* + o(|x_k - x^*|),$$
$$|\nabla f(x_k)| = O(|x_k - x^*|).$$

Using the above two relations in (33), we obtain

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

and (32) becomes

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

from which

$$\lim_{k \to \infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|} = \lim_{k \to \infty} \frac{o(|x_k - x^*|)}{|x_k - x^*|} = 0. \qquad \text{Q.E.D.}$$

We note that one can prove that Eq. (30) is equivalent to

$$\lim_{k \to \infty} \frac{|[D_k^{-1} - \nabla^2 f(x^*)] D_k \nabla f(x_k)|}{|D_k \nabla f(x_k)|} = 0 \tag{34}$$

assuming (28) holds. Equation (34) has been used by Dennis and Moré (1974) in the analysis of quasi-Newton methods and is sometimes called the *Dennis–Moré condition* (see also McCormick and Ritter, 1972).

A slight modification of the proof of Proposition 1.15 shows also that its conclusion holds if α_k is chosen by means of the Goldstein rule with initial trial stepsize equal to unity. Furthermore for all k sufficiently large, we shall have $\alpha_k = 1$ (i.e., the initial stepsize will be acceptable after a certain index).

Several additional results relating to the convergence rate of gradient methods are possible. The main guideline which consistently emerges from this analysis (and which has been supported by extensive numerical experience) is that *in order to achieve fast convergence of the iteration*

$$x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k),$$

one should try to choose the matrices D_k as close as possible to $[\nabla^2 f(x^*)]^{-1}$ so that the corresponding maximum and minimum eigenvalues of $D_k^{1/2}\nabla^2 f(x^*)D_k^{1/2}$ satisfy $M_k \sim 1$ and $m_k \sim 1$. This fact holds true for all stepsize rules that we have examined. Furthermore, when $M_k \sim 1$ and $m_k \sim 1$, the initial stepsize $s = 1$ is a good choice for the Armijo rule and other related rules or as a starting point for one-dimensional minimization procedures in minimization stepsize rules.

Spacer Steps in Descent Algorithms

Often in optimization problems, we utilize complex descent algorithms in which the rule used to determine the next point may depend on several previous points or on the iteration index k. Some of the conjugate direction algorithms to be examined in the next chapter are of this type. Other algorithms may represent a combination of different methods and switch from one method to the other in a manner which may either be prespecified or may depend on the progress of the algorithm. Such combinations are usually introduced in order to improve speed of convergence or reliability. However, their convergence analysis can become extremely complicated. It is thus often of value to know that if in such algorithms one inserts, perhaps irregularly but infinitely often, an iteration of a convergent algorithm such as steepest descent, then the theoretical convergence properties of the overall algorithm are quite satisfactory. Such an iteration will be referred to as a *spacer step*. The related convergence result is given in the following proposition. The only requirement imposed on the iterations of the algorithm other than the spacer steps is that they do not increase the value of the objective function.

Proposition 1.16: Consider a sequence $\{x_k\}$ such that

$$f(x_{k+1}) \leq f(x_k) \qquad \forall\, k = 0, 1, \ldots .$$

Assume that there exists an infinite set K of nonnegative integers for which we have

$$x_{k+1} = x_k + \alpha_k d_k \qquad \forall\, k \in K,$$

where $\{d_k\}_K$ is uniformly gradient related and α_k is chosen by the minimization rule, or the limited minimization rule, or the Armijo rule. Then every limit point of the subsequence $\{x_k\}_K$ is a critical point.

The proof requires a simple modification of the proof of Proposition 1.8 and is left to the reader. Notice that if f is a convex function, it is possible to strengthen the conclusion of the proposition and assert that every limit point of the whole sequence $\{x_k\}$ is a global minimum of f.

1.3.2 Steepest Descent and Scaling

Consider the steepest descent method

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k),$$

and assume that $f \in C^2$. We saw in the previous section that the convergence rate depends on the eigenvalue structure of the Hessian matrix $\nabla^2 f$. This structure in turn depends strongly on the particular choice of variables x used to define the problem. A different choice may change substantially the convergence rate.

Let T be an invertible $n \times n$ matrix. We can then represent points in R^n either by the vector x which enters in the objective function $f(x)$, or by the vector y, where

$$Ty = x. \tag{35}$$

Then the problem of minimizing f is equivalent to the problem

$$\begin{aligned} &\text{minimize} \quad h(y) \triangleq f(Ty) \\ &\text{subject to} \quad y \in R^n. \end{aligned} \tag{36}$$

If y^* is a local minimum of h, the vector $x^* = Ty^*$ is a local minimum of f.

Now steepest descent for problem (36) takes the form

$$y_{k+1} = y_k - \alpha_k \nabla h(y_k) = y_k - \alpha_k T' \nabla f(Ty_k). \tag{37}$$

Multiplying both sides by T and using (35) we obtain the iteration in terms of the x variables

$$x_{k+1} = x_k - \alpha_k TT' \nabla f(x_k).$$

Setting $D = TT'$, we obtain the following scaled version of steepest descent

$$x_{k+1} = x_k - \alpha_k D \nabla f(x_k) \tag{38}$$

with D being a positive definite symmetric matrix. The convergence rate of (37) or equivalently (38), however, is governed by the eigenvalue structure of $\nabla^2 h$ rather than of $\nabla^2 f$. We have $\nabla^2 h(y) = T' \nabla^2 f(Ty) T$, and if T is symmetric and positive definite, then $T = D^{1/2}$ and

$$\nabla^2 h(y) = D^{1/2} \nabla^2 f(x) D^{1/2}.$$

When $D \sim [\nabla^2 f(x)]^{-1}$, we obtain $\nabla^2 h(y) \sim I$, and the problem of minimizing h becomes well scaled and can be solved efficiently by steepest descent. This is consistent with the rate of convergence results of the previous section.

The more general iteration

$$x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k)$$

with D_k positive definite may be viewed as a scaled version of the steepest descent method where at each iteration we use different scaling for the variables. Good scaling is obtained when $D_k \sim [\nabla^2 f(x^*)]^{-1}$, where x^* is a local minimum to which the method is assumed to converge ultimately. Since $\nabla^2 f(x^*)$ is unavailable, often we use $D_k = [\nabla^2 f(x_k)]^{-1}$ or $D = [\nabla^2 f(x_0)]^{-1}$, where these matrices are positive definite. This type of scaling results in modified forms of Newton's method. A less complicated form of scaling is obtained when D is chosen to be diagonal of the form

$$D = \begin{bmatrix} d^1 & & & 0 \\ & d^2 & & \\ & & \ddots & \\ 0 & & & d^n \end{bmatrix}$$

with

$$d^i \sim [\partial^2 f(x_0)/(\partial x^i)^2]^{-1}, \qquad i = 1, \ldots, n;$$

i.e., the Hessian matrix is approximated by a diagonal matrix. The approximate inverse second derivatives d^i are obtained either analytically or by finite differences of first derivatives at the starting point x_0. It is also possible to update the scaling factors d^i periodically. The scaled version of steepest descent takes the form

$$x_{k+1}^i = x_k^i - \alpha_k d^i \, \partial f(x_k)/\partial x^i, \qquad i = 1, \ldots, n.$$

While such simple scaling schemes are not guaranteed to improve the convergence rate of steepest descent, in many cases they can result in spectacular improvements. An additional advantage when using the simple diagonal scaling device described above is that usually the initial stepsize $s = 1$ will work well for the Armijo rule, thus eliminating the need for determining a range of good initial stepsize choices by experimentation.

1.3.3 Newton's Method and Its Modifications

Newton's method consists of the iteration

$$x_{k+1} = x_k - \alpha_k [\nabla^2 f(x_k)]^{-1} \nabla f(x_k), \tag{39}$$

assuming that $[\nabla^2 f(x_k)]^{-1}$ exists and that the Newton direction

$$d_k = -[\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$$

is a direction of descent (i.e., $d_k' \nabla f(x_k) < 0$). This direction is obtained as the solution of the linear system of equations

$$\nabla^2 f(x_k) d_k = -\nabla f(x_k).$$

As explained in the section on scaling, one may view this iteration as a scaled version of steepest descent where the "optimal" scaling matrix $D_k = [\nabla^2 f(x_k)]^{-1}$ is utilized. It is also worth mentioning that *Newton's method is "scale-free"* in the sense that the method cannot be affected by a change in coordinate system as is the case with steepest descent (Section 1.3.2). Indeed if we consider a linear invertible transformation of variables $x = Ty$, then Newton's method in the space of the variables y is written as

$$y_{k+1} = y_k - \alpha_k [\nabla^2_{yy} f(Ty_k)]^{-1} \nabla_y f(Ty_k) = y_k - \alpha_k T^{-1} \nabla^2 f(Ty_k)^{-1} \nabla f(Ty_k),$$

and by applying T to both sides of this equation we recover (39).

When the Armijo rule is utilized with initial stepsize $s = 1$, then no reduction of the stepsize will be necessary near convergence to a strong minimum, as shown in Proposition 1.15. Thus, near convergence the method takes the form

$$x_{k+1} = x_k - [\nabla^2 f(x_k)]^{-1} \nabla f(x_k), \tag{40}$$

which will be referred to as the *pure form of Newton's method.* A valuable interpretation of this iteration is obtained by observing that x_{k+1} as given above minimizes the second-order Taylor's series expansion of f around x_k given by

$$\tilde{f}_k(x) = f(x_k) + \nabla f(x_k)'(x - x_k) + \tfrac{1}{2}(x - x_k)' \nabla^2 f(x_k)(x - x_k).$$

Indeed by setting the derivative of $\tilde{f}_k$ equal to zero, we obtain

$$\nabla \tilde{f}_k(x) = \nabla f(x_k) + \nabla^2 f(x_k)(x - x_k) = 0.$$

The solution of this equation is x_{k+1} as given by Eq. (40). It follows that when f is positive definite quadratic the pure form of Newton's method yields the unique minimum of f in a single iteration. Thus, one expects that iteration (40) will have a fast rate of convergence. This is substantiated by the following result which applies to Newton's method for solving systems of equations:

Proposition 1.17: Consider a mapping $g: R^n \to R^n$, and let $\varepsilon > 0$ and x^* be such that $g \in C^1$ on $S(x^*; \varepsilon)$, $g(x^*) = 0$, and $\nabla g(x^*)$ is invertible. Then there exists a $\delta > 0$ such that if $x_0 \in S(x^*; \delta)$, the sequence $\{x_k\}$ generated by the iteration

$$x_{k+1} = x_k - [\nabla g(x_k)']^{-1} g(x_k)$$

is well defined, converges to x^*, and satisfies $x_k \in S(x^*; \delta)$ for all k. Furthermore, if $x_k \neq x^*$ for all k, then

$$\lim_{k \to \infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|} = 0; \tag{41}$$

i.e., $\{|x_k - x^*|\}$ converges Q-superlinearly. In addition given any $r > 0$, there exists a $\delta_r > 0$ such that if $x_k \in S(x^*; \delta_r)$, then

$$|x_{k+1} - x^*| \leq r|x_k - x^*|, \tag{42}$$

$$|g(x_{k+1})| \leq r|g(x_k)|. \tag{43}$$

If we assume further that for some $L > 0$ and $M > 0$, we have

$$|\nabla g(x)' - \nabla g(y)'| \leq L|x - y| \qquad \forall\, x, y \in S(x^*; \varepsilon), \tag{44a}$$

$$|[\nabla g(x)']^{-1}| \leq M, \qquad \forall\, x \in S(x^*; \varepsilon), \tag{44b}$$

then

$$|x_{k+1} - x^*| \leq \tfrac{1}{2}LM|x_k - x^*|^2 \qquad \forall\, k = 0, 1, \ldots,$$

and $\{|x_k - x^*|\}$ converges Q-superlinearly with order at least two.

Proof: Let $\delta \in (0, \varepsilon)$ and $M > 0$ be such that $[\nabla g(x)]^{-1}$ exists for all $x \in S(x^*; \delta)$ and

$$|[\nabla g(x)']^{-1}| \leq M \qquad \forall\, x \in S(x^*; \delta). \tag{45}$$

If $x_k \in S(x^*; \delta)$, we have

$$g(x_k) = \int_0^1 \nabla g[x^* + t(x_k - x^*)]'\, dt(x_k - x^*)$$

from which

$$\begin{aligned} x_{k+1} - x^* &= x_k - x^* - [\nabla g(x_k)']^{-1} g(x_k) \\ &= [\nabla g(x_k)']^{-1}[\nabla g(x_k)'(x_k - x^*) - g(x_k)] \\ &= [\nabla g(x_k)']^{-1}\left[\nabla g(x_k)' - \int_0^1 \nabla g[x^* + t(x_k - x^*)]'\, dt\right](x_k - x^*) \\ &= [\nabla g(x_k)']^{-1} \int_0^1 \{\nabla g(x_k)' - \nabla g[x^* + t(x_k - x^*)]'\}\, dt(x_k - x^*). \end{aligned} \tag{46}$$

By continuity of ∇g, we can take δ sufficiently small to ensure that

$$|\nabla g(x)' - \nabla g(y)'| < \tfrac{1}{2}M^{-1} \qquad \forall\, x, y \in S(x^*; \delta). \tag{47}$$

Then from (45), (46), and (47), we obtain

$$|x_{k+1} - x^*| \leq |[\nabla g(x_k)']^{-1}| \int_0^1 |\nabla g(x_k)' - \nabla g[x^* + t(x_k - x^*)]'|\, dt\, |x_k - x^*| \tag{48}$$

and

$$|x_{k+1} - x^*| < \tfrac{1}{2}|x_k - x^*|.$$

It follows that if $x_0 \in S(x^*; \delta)$ then $x_k \in S(x^*; \delta)$ for all k and $x_k \to x^*$. Equation (41) then follows from (48).

We have

$$g_i(x) = \nabla g_i(\tilde{x}_i)'(x - x^*) \qquad \forall\, i = 1, \ldots, n,$$

where $\tilde{x}_i$ is a vector lying in the line segment connecting x and x^*. Therefore by denoting $\nabla g(\tilde{x})$ the matrix with columns $\nabla g_i(\tilde{x}_i)$, we have

$$|g(x)|^2 = (x - x^*)'\nabla g(\tilde{x})\nabla g(\tilde{x})'(x - x^*).$$

Choose $\delta_1 > 0$ sufficiently small so that $\nabla g(\tilde{x})\nabla g(\tilde{x})'$ is positive definite for all x with $|x - x^*| \leq \delta_1$, and let $\Lambda > 0$ and $\lambda > 0$ be upper and lower bounds to the eigenvalues of $[\nabla g(\tilde{x})\nabla g(\tilde{x})']^{1/2}$ for $x \in S(x^*; \delta_1)$. Then

$$\begin{aligned}\lambda^2|x - x^*|^2 &\leq (x - x^*)'\nabla g(\tilde{x})\nabla g(\tilde{x})'(x - x^*)\\ &\leq \Lambda^2|x - x^*|^2 \qquad \forall\, x \in S(x^*; \delta_1).\end{aligned}$$

Hence, we have

$$\lambda|x - x^*| \leq |g(x)| \leq \Lambda|x - x^*| \qquad \forall\, x \in S(x^*; \delta_1).$$

Now from (48), it follows easily that given any $r > 0$, we can find a $\delta_r \in (0, \delta_1]$ such that if $x_k \in S(x^*; \delta_r)$, then

$$|x_{k+1} - x^*| \leq (\lambda r/\Lambda)|x_k - x^*| \leq r|x_k - x^*|,$$

thereby showing (42). Combining the last two inequalities we also obtain

$$|g(x_{k+1})| \leq r|g(x_k)| \qquad \forall\, x \in S(x^*; \delta_r),$$

and (43) is proved.

If (44a) and (44b) hold, then from (48) we have

$$|x_{k+1} - x^*| \leq M \int_0^1 Lt|x_k - x^*|\, dt|x_k - x^*| = \frac{ML}{2}|x_k - x^*|^2.$$

Q.E.D.

For $g(x) = \nabla f(x)$, the result of the proposition applies to the pure form of Newton's method (40). Extensive computational experience suggests that the fast convergence rate indicated in the proposition is indeed realized in a practical setting. On the other hand, Newton's method in its pure form has several serious drawbacks. First, the inverse $[\nabla^2 f(x_k)]^{-1}$ may fail to exist, in which case the method breaks down. This may happen, for example, if f is linear within some region in which case $\nabla^2 f = 0$. Second, iteration (40) is not a descent method in the sense that it may easily happen that $f(x_{k+1}) > f(x_k)$. Third, the method tends to be attracted by local maxima just as much

as it is attracted by local minima. This is evident from Proposition 1.17 where it is assumed that $\nabla g(x^*)$ is invertible but not necessarily positive definite.

For these reasons, it is necessary to modify the pure form of Newton's method (40) in order to convert it to a reliable minimization algorithm. There are several schemes by means of which this can be accomplished. All these schemes convert iteration (40) into a gradient method with a uniformly gradient-related direction sequence, while guaranteeing that whenever the algorithm gets sufficiently close to a point x^* satisfying the second-order sufficiency conditions, then the algorithm assumes the pure form (40) and achieves the attendant fast convergence rate.

First Modification Scheme: This method consists of the iteration

$$x_{k+1} = x_k + \alpha_k d_k,$$

where α_k is chosen by the Armijo rule with initial stepsize unity ($s = 1$), and d_k is chosen by

$$d_k = -[\nabla^2 f(x_k)]^{-1} \nabla f(x_k), \tag{49}$$

if $[\nabla^2 f(x_k)]^{-1}$ exists and

$$\nabla f(x_k)'[\nabla^2 f(x_k)]^{-1} \nabla f(x_k) \geq c_1 |\nabla f(x_k)|^{p_1}, \tag{50}$$

$$c_2 |\nabla f(x_k)| \geq |[\nabla^2 f(x_k)]^{-1} \nabla f(x_k)|^{p_2}, \tag{51}$$

while otherwise

$$d_k = -D \nabla f(x_k).$$

The matrix D is some positive definite symmetric scaling matrix. The scalars c_1, c_2, p_1, and p_2 satisfy

$$c_1 > 0, \qquad c_2 > 0, \qquad p_1 > 2, \qquad \text{and} \qquad p_2 > 1.$$

In practice c_1 should be very small, say 10^{-5}, c_2 should be very large, say 10^5, and p_1 and p_2 can be chosen equal to three and two, respectively.

It is clear, from Proposition 1.8, that a sequence $\{d_k\}$ generated by the scheme above is uniformly gradient related and hence the resulting algorithm is convergent in the sense that every limit point of a sequence that it generates is a critical point of f. Now consider the algorithm near a local minimum x^* satisfying

$$\nabla f(x^*) = 0, \qquad \nabla^2 f(x^*) > 0.$$

Then it is easy to see that for x_k close enough to x^*, the Hessian $\nabla^2 f(x_k)$ will be invertible and the tests (50) and (51) will be passed. Thus, d_k will be the Newton direction (49) for all x_k sufficiently close to x. Furthermore, from Propositions 1.12 and 1.15, we shall have $x_k \to x^*$, and the stepsize

α_k will equal unity. Hence, if x_k is sufficiently close to x^*, then $x_k \to x^*$, and the pure form of Newton's method will be employed after some index, thus achieving the fast convergence rate indicated in Proposition 1.17.

A variation of this modification scheme is given by the iteration

$$x_{k+1} = x_k + \alpha_k[\alpha_k d_k^N - (1 - \alpha_k)D\nabla f(x_k)],$$

where D is a positive definite matrix and d_k^N is the Newton direction

$$d_k^N = -[\nabla^2 f(x_k)]^{-1}\nabla f(x_k),$$

if $[\nabla^2 f(x_k)]^{-1}$ exists. Otherwise $d_k^N = -D\nabla f(x_k)$. The stepsize α_k is chosen by an Armijo-type rule with initial stepsize unity whereby $\alpha_k = \beta^{m_k}$ and m_k is the first nonnegative integer m for which

$$f(x_k) - f[x_k + \beta^m d_k(\beta^m)] \geq -\sigma\beta^m |\nabla f(x_k)|^2,$$

where $\sigma \in (0, \frac{1}{2})$, $\beta \in (0, 1)$, and

$$d_k(\beta^m) = \beta^m d_k^N - (1 - \beta^m)D\nabla f(x_k).$$

This is a line search along the curve of points of the form

$$z_\alpha = \alpha[\alpha d_k^N - (1 - \alpha)D\nabla f(x_k)]$$

with $\alpha \in [0, 1]$. For $\alpha = 1$ we obtain the Newton direction, while as $\alpha \to 0$ the vector z_α/α tends to the (scaled) steepest descent direction $-D\nabla f(x_k)$. Assuming σ is chosen sufficiently small, one can prove similar convergence and rate of convergence results as the ones stated earlier for this modified version of Newton's method.

Second Modification Scheme: Since calculation of the Newton direction d_k involves solution of the system of linear equations

$$\nabla^2 f(x_k)d_k = -\nabla f(x_k),$$

it is natural to compute d_k by attempting to form the Cholesky factorization of $\nabla^2 f(x_k)$ (see the preceding section). During the factorization process, one can detect whether $\nabla^2 f(x_k)$ is either nonpositive definite or nearly singular, in which case $\nabla^2 f(x_k)$ is replaced by a positive definite matrix of the form $F_k = \nabla^2 f(x_k) + E_k$, where E_k is a diagonal matrix. The elements of E_k are introduced sequentially during the factorization process, so that at the end we obtain F_k in the form $F_k = L_k L_k'$, where L_k is lower triangular. Subsequently d_k is obtained as the solution of the system of equations $L_k L_k' d_k = -\nabla f(x_k)$, and the next point x_{k+1} is determined from $x_{k+1} = x_k + \alpha_k d_k$, where α_k is chosen according to the Armijo rule. The matrix E_k is such that the sequence $\{d_k\}$ is uniformly gradient related. Furthermore, $E_k = 0$ when x_k is close enough to a point x^* satisfying the second-order

sufficiency conditions for optimality. Thus, near such a point, the method is again identical to the pure form of Newton's method and achieves the corresponding superlinear convergence rate. The precise mechanization of the scheme is as follows.

Let $c > 0$, $\mu > 0$, and $p > 0$ denote fixed scalars and let a_{ij}^k denote the elements of $\nabla^2 f(x_k)$. Consider the $i \times i$ lower triangular matrices L_k^i, $i = 1, \ldots, n$, defined recursively by the following modified Cholesky factorization process (compare with Section 1.2):

$$L_k^1 = \begin{cases} \sqrt{a_{11}^k} & \text{if } a_{11}^k > 0 \quad \text{and} \quad \sqrt{a_{11}^k} \geq c|\nabla f(x_k)|^p, \\ \mu & \text{otherwise,} \end{cases}$$

$$L_k^i = \begin{bmatrix} L_k^{i-1} & 0 \\ l_i^{k\prime} & \lambda_{ii}^k \end{bmatrix}, \qquad i = 2, \ldots, n,$$

where

$$l_i^k = (L_k^{i-1})^{-1} a_i^k, \qquad a_i^k = \begin{bmatrix} a_{1i}^k \\ \vdots \\ a_{i-1,i}^k \end{bmatrix},$$

$$\lambda_{ii}^k = \begin{cases} \sqrt{a_{ii}^k - l_i^{k\prime} l_i^k} & \text{if } a_{ii}^k > l_i^{k\prime} l_i^k \quad \text{and} \quad \sqrt{a_{ii}^k - l_i^{k\prime} l_i^k} \geq c|\nabla f(x_k)|^p, \\ \mu & \text{otherwise.} \end{cases}$$

Then the direction d_k is determined from

$$L_k L_k' d_k = -\nabla f(x_k),$$

where $L_k = L_k^n$. The next point x_{k+1} is determined from

$$x_{k+1} = x_k + \alpha_k d_k,$$

where α_k is chosen by the Armijo rule with initial stepsize $s = 1$ whenever $\nabla^2 f(x_k) = L_k L_k'$.

Some trial and error may be necessary in order to determine appropriate values for c, μ, and p. Usually, one takes c very small so that the Newton direction will be modified as infrequently as possible. The value of μ should be considerably larger than that of c in order that the matrix $L_k L_k'$ is not nearly singular. A choice $0 < p \leq 1$ is usually satisfactory. Sometimes one takes $p = 0$, although in this case the theoretical convergence rate properties of the algorithm depend on the value of c.

The following facts may be verified for the algorithm described above:

(a) The direction sequence $\{d_k\}$ is uniformly gradient related, and hence the resulting algorithm is convergent in the sense that every limit point of $\{x_k\}$ is a critical point of f.

(b) For each point x^* satisfying $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) > 0$, there exists a scalar $\varepsilon > 0$ such that if $|x_k - x^*| < \varepsilon$ then $L_k L_k' = \nabla^2 f(x_k)$; i.e., the Newton direction will not be modified, and furthermore the stepsize α_k will equal unity. Thus, when sufficiently close to such a point x^*, the algorithm assumes the pure form of Newton's method and converges to x^* with superlinear convergence rate.

There is another interesting modification scheme that can be used when $\nabla^2 f(x_k)$ is indefinite. In this case one can use, instead of the direction $[\nabla^2 f(x_k)]^{-1} \nabla f(x_k)$, a descent direction which is also *a direction of negative curvature*, i.e., a d_k such that $\nabla f(x_k)' d_k < 0$ and $d_k' \nabla^2 f(x_k) d_k < 0$. This can be done in a numerically stable and efficient manner via a form of triangular factorization of $\nabla^2 f(x_k)$. For a detailed presentation we refer to Fletcher and Freeman (1977), More and Sorensen (1979), and Goldfarb (1980).

Periodic Reevaluation of the Hessian

Finally, we mention that a Newton-type method, which in many cases is considerably more efficient computationally than those described above, is obtained if the Hessian matrix $\nabla^2 f$ is recomputed every p iterations ($p \geq 2$) rather than at every iteration. This method in unmodified form is given by

$$x_{k+1} = x_k - \alpha_k D_k \nabla f(x_k),$$

where

$$D_{ip+j} = [\nabla^2 f(x_{ip})]^{-1}, \qquad j = 0, 1, \ldots, p-1, \quad i = 0, 1, \ldots.$$

A significant advantage of this method when coupled with the second modification scheme described above is that the Cholesky factorization of $\nabla^2 f(x_{ip})$ is obtained at the ipth iteration and is subsequently used for a total of p iterations in the computation of the direction of search. This reduction in computational burden per iteration is achieved at the expense of what is usually a small or imperceptible degradation in speed of convergence.

Approximate Newton Methods

One of the main drawbacks of Newton's method in its pure or modified forms is the need to solve a system of linear equations in order to obtain the descent direction at each iteration. We have so far implicitly assumed that this system will be solved by some version of the Gaussian elimination method which requires a finite number of arithmetic operations [$O(n^3)$]. On the other hand, if the dimension n is large, the amount of calculation required for exact solution of the Newton system can be prohibitive and one may have to be satisfied with only an approximate solution of this system. This approach is often used in fact for solving large linear systems

of equations where in some cases an adequate approximation to the solution can be obtained by iterative methods such as successive overrelaxation (SOR) much faster than the exact solution can be obtained by Gaussian elimination. The fact that Gaussian elimination can solve the system in a finite number of arithmetic operations while this is not guaranteed by SOR methods can be quite irrelevant, since the computational cost of finding the exact solution can be entirely prohibitive.

Another possibility is to solve the Newton system approximately by using the conjugate gradient method to be presented in the next section. More generally any system of the form $Hd = -g$, where H is a positive definite symmetric $n \times n$ matrix and $g \in R^n$ can be solved by the conjugate gradient method by converting it to the quadratic optimization problem

$$\begin{aligned} &\text{minimize} \quad \tfrac{1}{2}d'Hd + g'd \\ &\text{subject to} \quad d \in R^n. \end{aligned}$$

It will be seen in the next section that actually the conjugate gradient method solves this problem exactly in at most n iterations. However this fact is not particularly relevant since for the type of problems where the use of the conjugate gradient method makes sense, the dimension n is very large and the main hope is that only a few conjugate gradient steps will be necessary in order to obtain a good approximation to the solution.

For the purposes of unconstrained optimization, an important property of any approximate method of solving a system of the form $H_k d = -\nabla f(x_k)$, where H_k is positive definite, is that the approximate direction $\bar{d}$ obtained is a descent direction, i.e., it satisfies $\nabla f(x_k)'\bar{d} < 0$. This will be automatically satisfied if the approximate method used is a descent method for solving the quadratic optimization problem

$$\begin{aligned} &\text{minimize} \quad \tfrac{1}{2}d'H_k d + \nabla f(x_k)'d \\ &\text{subject to} \quad d \in R^n, \end{aligned}$$

and the starting point $d_0 = 0$ is used, for the descent property implies

$$\tfrac{1}{2}\bar{d}'H_k\bar{d} + \nabla f(x_k)'\bar{d} < \tfrac{1}{2}d_0'H_k d_0 + \nabla f(x_k)'d_0 = 0,$$

or $\nabla f(x_k)'\bar{d} < -\tfrac{1}{2}\bar{d}'H_k\bar{d} < 0$. As will be seen in the next section, the conjugate gradient method has this property.

Conditions on the accuracy of the approximate solution $\bar{d}$ that ensure linear or superlinear rate of convergence in connection with approximate methods are given in Dembo *et al.* (1980). Generally speaking if $H_k \to \nabla^2 f(x_k)$ and the approximate Newton directions d_k satisfy

$$\lim_{k\to\infty} \frac{|H_k d_k + \nabla f(x_k)|}{|\nabla f(x_k)|} \to 0,$$

the superlinear convergence rate property of the method to a strong local minimum is maintained (compare with Proposition 1.15). Approximate Newton methods based on the conjugate gradient method are applied to large scale nonlinear multicommodity flow problems in Bertsekas and Gafni (1981).

1.3.4 Conjugate Direction and Conjugate Gradient Methods

Conjugate direction methods are motivated by a desire to accelerate the convergence rate of steepest descent while avoiding the overhead and evaluation of second derivatives associated with Newton's method. Conjugate direction methods are typically analyzed for the purely quadratic problem

$$\text{(52)} \qquad \begin{aligned} &\text{minimize} \quad f(x) = \tfrac{1}{2}x'Qx \\ &\text{subject to} \quad x \in R^n, \end{aligned}$$

where $Q > 0$, which they can solve in at most n iterations (see Proposition 1.18 that follows). It is then argued that the general problem can be approximated near a strong local minimum by a quadratic problem. One therefore expects that conjugate direction methods, suitably modified, should work well for the general problem—a conjecture that has been substantiated by analysis as well as practical experience.

Definition: Given a positive definite $n \times n$ matrix Q, we say that a collection of nonzero vectors $d_1, \ldots, d_k \in R^n$ is mutually *Q-conjugate* if for all i and j with $i \neq j$ we have $d_i'Qd_j = 0$.

It is clear that *if $d_1, \ldots, d_k$ are mutually Q-conjugate then they are linearly independent*, since if, for example, we had for scalars $\alpha_1, \ldots, \alpha_{k-1}$

$$d_k = \alpha_1 d_1 + \cdots + \alpha_{k-1}d_{k-1},$$

then

$$d_k'Qd_k = \alpha_1 d_k'Qd_1 + \cdots + \alpha_{k-1}d_k'Qd_{k-1} = 0,$$

which is impossible since $d_k \neq 0$, and Q is positive definite.

Given a collection of mutually Q-conjugate directions $d_0, \ldots, d_{n-1}$, we define the corresponding *conjugate direction method* for solving problem (52) by

$$\text{(53)} \qquad x_{k+1} = x_k + \alpha_k d_k, \qquad k = 0, 1, \ldots, n-1,$$

where x_0 is a given vector in R^n and α_k is defined by the line minimization rule

$$\text{(54)} \qquad f(x_k + \alpha_k d_k) = \min_{\alpha} f(x_k + \alpha d_k).$$

We shall employ in what follows in this and the next section the notation

$$g_k = \nabla f(x_k) = Qx_k.$$

We have the following result:

Proposition 1.18: If $x_1, x_2, \ldots, x_n$ are the vectors generated by the conjugate direction method (53), we have

$$g'_{k+1}d_i = 0 \qquad \forall\, i = 0, \ldots, k. \tag{55}$$

Furthermore, for $k = 0, 1, \ldots, n-1$, x_{k+1} minimizes f over the linear manifold

$$M_k = \{z \mid z = x_0 + \gamma_0 d_0 + \cdots + \gamma_k d_k, \gamma_0, \ldots, \gamma_k \in R\},$$

and hence x_n minimizes f over R^n.

Proof: By (54), we have

$$\partial f(x_i + \alpha_i d_i)/\partial\alpha = g'_{i+1}d_i = 0, \qquad i = 0, \ldots, n-1,$$

so we need only verify (55) for $i = 0, 1, \ldots, k-1$. We have, for $i = 0, 1, \ldots, k-1$,

$$g'_{k+1}d_i = x'_{k+1}Qd_i = \left(x_{i+1} + \sum_{j=i+1}^{k} \alpha_j d_j\right)' Qd_i = x'_{i+1}Qd_i = g'_{i+1}d_i = 0.$$

To show the last part of the proposition, we must show that

$$\partial f(x_0 + \gamma_0 d_0 + \cdots + \gamma_k d_k)/\partial\gamma_i|_{\substack{\gamma_0 = \alpha_0 \\ \vdots \\ \gamma_k = \alpha_k}} = 0 \qquad \forall\, i = 0, \ldots, k$$

or

$$g'_{k+1}d_i = 0 \qquad \forall\, i = 0, \ldots, k,$$

which is (55). Q.E.D.

It is easy to visualize the result of Proposition 1.18 for the case where $Q = I$, for in this case, the surfaces of equal cost of f are concentric spheres, and the notion of Q-conjugacy reduces to usual orthogonality. By elementary geometry or a simple algebraic argument, we have that minimization along n orthogonal directions leads to the global minimum of f, i.e., the center of the spheres. The case of a general positive definite Q can actually be reduced to the case where $Q = I$ by means of a scaling transformation. By setting $y = Q^{1/2}x$, the problem becomes min $\{\frac{1}{2}|y|^2 \mid y \in R^n\}$. If $w_0, \ldots, w_{n-1}$ are any set of orthogonal nonzero vectors in R^n, the algorithm

$$y_{k+1} = y_k + \alpha_k w_k, \qquad k = 0, 1, \ldots, n-1,$$

where α_k minimizes $\frac{1}{2}|y_k + \alpha w_k|^2$ over α, terminates in at most n steps at $y_n = 0$. To pass back to the x-coordinate system, we multiply this equation by $Q^{-1/2}$ and obtain

$$x_{k+1} = x_k + \alpha_k d_k, \qquad k = 0, 1, \ldots, n-1,$$

where $d_k = Q^{-1/2} w_k$. Since $w_i' w_j = 0$ for $i \neq j$, we obtain $d_i' Q d_j = 0$ for $i \neq j$; i.e., the directions $d_0, \ldots, d_{n-1}$ are Q-conjugate. This argument can be reversed and shows that the collection of conjugate direction methods for the problem $\min\{\frac{1}{2}x'Qx \mid x \in R^n\}$ is in one-to-one correspondence with the set of methods for solving the problem $\min\{\frac{1}{2}|y|^2 \mid y \in R^n\}$, which consist of successive minimization along n orthogonal directions.

Given any set of linearly independent vectors $\xi_0, \ldots, \xi_{n-1}$, we can construct a set of mutually Q-conjugate directions $d_0, \ldots, d_{n-1}$ as follows. Set

$$d_0 = \xi_0, \tag{56}$$

and for $i = 1, 2, \ldots, n-1$, define successively

$$d_i = \xi_i + \sum_{j=0}^{i-1} c_{ij} d_j, \tag{57}$$

where the coefficients c_{ij} are chosen so that d_i is Q-conjugate to the previous directions $d_{i-1}, \ldots, d_0$. This will be so if, for $k = 0, \ldots, i-1$,

$$d_i' Q d_k = \xi_i' Q d_k + \sum_{j=0}^{i-1} c_{ij} d_j' Q d_k = 0. \tag{58}$$

If previous coefficients were chosen so that $d_0, \ldots, d_{i-1}$ are Q-conjugate, then we have $d_j' Q d_k = 0$ if $j \neq k$, and (58) yields

$$c_{ij} = -\xi_i' Q d_j / d_j' Q d_j \qquad \forall\, i = 1, 2, \ldots, n-1, \quad j = 0, 1, \ldots, i-1. \tag{59}$$

Thus the set of directions $d_0, \ldots, d_{n-1}$ defined by (56), (57) and (59) is Q-conjugate, and (56) and (57) show also that, for $i = 0, \ldots, n-1$, we have

$$\text{(subspace spanned by } d_0, \ldots, d_i) = \text{(subspace spanned by } \xi_0, \ldots, \xi_i). \tag{60}$$

We now define the most important conjugate direction method.

The Conjugate Gradient Method

The conjugate gradient method is obtained by the procedure described above by taking $\xi_0 = -g_0, \ldots, \xi_{n-1} = -g_{n-1}$. More specifically, starting at x_0 with $g_0 \neq 0$, we use g_0 as our first conjugate direction, i.e., $d_0 = -g_0$. We find $x_1 = x_0 + \alpha_0 d_0$ by line search and obtain our second direction

d_1 using the procedure defined by (56), (57) and (59) with $\xi_0 = -g_0$ and $\xi_1 = -g_1$. This yields, from (57) and (59),

$$d_1 = -g_1 + \frac{g_1' Q d_0}{d_0' Q d_0} d_0. \tag{61}$$

By using the equation

$$g_1 - g_0 = Q(x_1 - x_0) = \alpha_0 Q d_0,$$

we can write (61) as

$$d_1 = -g_1 + \frac{g_1'(g_1 - g_0)}{d_0'(g_1 - g_0)} d_0.$$

By repeating the process with $\xi_0 = -g_0$, $\xi_1 = -g_1, \ldots,$ and $\xi_k = -g_k$, we obtain at the $(k+1)$st step

$$d_k = -g_k + \sum_{j=0}^{k-1} \frac{g_k' Q d_j}{d_j' Q d_j} d_j$$

from which

$$d_k = -g_k + \sum_{j=0}^{k-1} \frac{g_k'(g_{j+1} - g_j)}{d_j'(g_{j+1} - g_j)} d_j. \tag{62}$$

By using the fact that the subspace spanned by $g_0, \ldots, g_{k-1}$ is also the subspace spanned by $d_0, \ldots, d_{k-1}$ [compare with (60)] and the relation $g_k' d_j = 0$ for $j = 0, \ldots, k-1$ (Proposition 1.18), we obtain

$$g_k' g_j = 0, \qquad j = 0, \ldots, k-1,$$

so (62) reduces to the simple formula

$$d_k = -g_k + \beta_k d_{k-1}, \tag{63}$$

with

$$\beta_k = \frac{g_k'(g_k - g_{k-1})}{d_{k-1}'(g_k - g_{k-1})}. \tag{64}$$

Note that by using the facts $g_k' g_j = g_k' d_j = 0, j = 0, \ldots, k-1$, and $d_{k-1} = -g_{k-1} + \beta_{k-1} d_{k-2}$, we see that the coefficient β_k of (64) can also be written as

$$\beta_k = \frac{g_k'(g_k - g_{k-1})}{g_{k-1}' g_{k-1}} = \frac{g_k' g_k}{g_{k-1}' g_{k-1}}. \tag{65}$$

An important observation from (63) and (64) is that *in order to generate the direction d_k one need only know the current and previous gradients g_k and*

g_{k-1} *and the previous direction* d_{k-1}. This fact is particularly significant when the method is extended to nonquadratic problems.

Scaled Conjugate Gradient Method

This method, also referred to as the *preconditioned conjugate gradient method*, is really the conjugate gradient method implemented in a new coordinate system. Suppose we make a change of variables, as in Section 1.3.2, $x = Ty$, where T is a symmetric invertible $n \times n$ matrix, and apply the conjugate gradient method to the equivalent problem

$$\begin{aligned} &\text{minimize} \quad h(y) = f(Ty) = \tfrac{1}{2} y'TQTy \\ &\text{subject to} \quad y \in R^n. \end{aligned}$$

The method is described by [compare with (63) and (65)]

$$y_{k+1} = y_k + \alpha_k \tilde{d}_k, \tag{66}$$

where α_k is obtained by line minimization and $\tilde{d}_k$ is generated by

$$\tilde{d}_0 = -\nabla h(y_0), \qquad \tilde{d}_k = -\nabla h(y_k) + \beta_k \tilde{d}_{k-1}, \qquad k = 1, 2, \ldots, n, \tag{67}$$

where

$$\beta_k = \frac{\nabla h(y_k)'\nabla h(y_k)}{\nabla h(y_{k-1})'\nabla h(y_{k-1})}. \tag{68}$$

Setting $x_k = Ty_k$, $\nabla h(y_k) = Tg_k$, $d_k = T\tilde{d}_k$, and $H = T^2$, we obtain from (66)–(68) the equivalent method

$$x_{k+1} = x_k + \alpha_k d_k, \tag{69}$$

$$d_0 = -Hg_0, \qquad d_k = -Hg_k + \beta_k d_{k-1}, \qquad k = 1, \ldots, n, \tag{70}$$

where

$$\beta_k = g_k' H g_k / g_{k-1}' H g_{k-1}. \tag{71}$$

Since $\nabla^2 h(y) = TQT$, we have that $\tilde{d}_0, \ldots, \tilde{d}_{n-1}$ are (TQT)-conjugate, and in view of $d_k = T\tilde{d}_k$, we have that $d_0, \ldots, d_{n-1}$ are Q-conjugate. By carrying further this line of argument we see that

$$g_k' H g_j = g_k' d_j = 0 \qquad \forall\, j = 0, \ldots, k-1,$$

and x_k minimizes f over the linear manifold

$$\begin{aligned} M_k &= \{z \mid z = x_0 + \gamma_0 d_0 + \cdots + \gamma_{k-1} d_{k-1}, \gamma_0, \ldots, \gamma_{k-1} \in R\} \\ &= \{z \mid z = x_0 + \gamma_0 H g_0 + \cdots + \gamma_{k-1} H g_{k-1}, \gamma_0, \ldots, \gamma_{k-1} \in R\}. \end{aligned}$$

The motivation for employing scaling typically stems from a desire to improve the speed of convergence of the method within an n-iteration cycle (see the following analysis). This in turn may be important even for a quadratic problem if n is large.

Rate of Convergence of the Conjugate Gradient Method

There are a number of results relating to the convergence rate of the conjugate gradient method applied to quadratic problems. We describe a particular result due to Luenberger (1973).

Consider an algorithm of the form

$$\begin{aligned} x_1 &= x_0 + \gamma_{00} g_0, \\ x_2 &= x_0 + \gamma_{10} g_0 + \gamma_{11} g_1, \\ &\vdots \\ x_{k+1} &= x_0 + \gamma_{k0} g_0 + \cdots + \gamma_{kk} g_k, \end{aligned} \tag{72}$$

where γ_{ij} are arbitrary scalars. Since $g_i = Qx_i$, we have that for suitable scalars ζ_{ki} the algorithm above can be written for all k

$$\begin{aligned} x_{k+1} &= x_0 + \zeta_{k0} Q x_0 + \zeta_{k1} Q^2 x_0 + \cdots + \zeta_{kk} Q^{k+1} x_0 \\ &= [I + QP_k(Q)]x_0, \end{aligned}$$

where P_k is a polynomial of degree k. Among all algorithms of the form (72), the conjugate gradient method is optimal in the sense that for every k, it minimizes $f(x_{k+1})$ over all sets of coefficients $\gamma_{k0}, \ldots, \gamma_{kk}$. It follows from the equation above that in the conjugate gradient method we have, for every k,

$$f(x_{k+1}) = \min_{P_k} \tfrac{1}{2} x_0' Q[I + QP_k(Q)]^2 x_0. \tag{73}$$

Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of Q, and let $e_1, \ldots, e_n$ be corresponding orthogonal eigenvectors normalized so that $|e_i| = 1$. Since $e_1, \ldots, e_n$ form a basis, any vector $x_0 \in R^n$ can be written as

$$x_0 = \sum_{i=1}^{n} \zeta_i e_i$$

for some scalars ζ_i. Since

$$Qx_0 = \sum_{i=1}^{n} \zeta_i Q e_i = \sum_{i=1}^{n} \zeta_i \lambda_i e_i,$$

we have, using the orthogonality of $e_1, \ldots, e_n$ and the fact that $|e_i| = 1$,

$$f(x_0) = \frac{1}{2} x_0' Q x_0 = \frac{1}{2} \left(\sum_{i=1}^{n} \zeta_i e_i \right)' \left(\sum_{i=1}^{n} \zeta_i \lambda_i e_i \right) = \frac{1}{2} \sum_{i=1}^{n} \lambda_i \zeta_i^2.$$

Applying the same process to (73), we obtain for *any* polynomial P_k of degree k

$$f(x_{k+1}) \leq \frac{1}{2}\sum_{i=1}^{n}[1 + \lambda_i P_k(\lambda_i)]^2 \lambda_i \zeta_i^2,$$

and it follows that

$$f(x_{k+1}) \leq \max_i [1 + \lambda_i P_k(\lambda_i)]^2 f(x_0) \qquad \forall P_k, k. \tag{74}$$

One can use this relationship for different choices of polynomials P_k to obtain a number of convergence rate results. We provide one such result.

Proposition 1.19: Assume that Q has $n - k$ eigenvalues in an interval $[a, b]$ with $a > 0$, and the remaining k eigenvalues are greater than b. Then for every x_0, the vector x_{k+1} generated after $(k + 1)$ steps of the conjugate gradient method satisfies

$$f(x_{k+1}) \leq \left(\frac{b - a}{b + a}\right)^2 f(x_0). \tag{75}$$

This relation also holds for the scaled conjugate gradient method (69)–(71) if the eigenvalues of Q are replaced by those of $H^{1/2}QH^{1/2}$.

Proof: Let $\lambda_1, \lambda_2, \ldots, \lambda_k$ be the eigenvalues of Q that are greater than b and consider the polynomial P_k defined by

$$1 + \lambda P_k(\lambda) = \frac{2}{(a + b)\lambda_1 \cdots \lambda_k}\left(\frac{a + b}{2} - \lambda\right)(\lambda_1 - \lambda)\cdots(\lambda_k - \lambda). \tag{76}$$

Since $1 + \lambda_i P_k(\lambda_i) = 0$ we have, using (74), (76), and a simple calculation,

$$\begin{aligned} f(x_{k+1}) &\leq \max_{a \leq \lambda \leq b} [1 + \lambda P_k(\lambda)]^2 f(x_0) \\ &\leq \max_{a \leq \lambda \leq b} \frac{[\lambda - \frac{1}{2}(a + b)]^2}{[\frac{1}{2}(a + b)]^2} f(x_0) = \left(\frac{b - a}{b + a}\right)^2 f(x_0). \end{aligned}$$

Q.E.D.

An immediate consequence of the proposition is that if the eigenvalues of Q take only k distinct values then the conjugate gradient method will find the minimum of the quadratic function f in at most k iterations. (Simply take $a = b$ in the proposition.) Another interesting possibility, arising for example in some optimal control problems, is when Q has the form

$$Q = M + \sum_{i=1}^{k} v_i v_i', \tag{77}$$

where M is positive definite symmetric, and v_i are some vectors in R^n. We have the following result, the proof of which we leave as an exercise for the reader.

Exercise: Show that if Q is of the form (77), then the vector x_{k+1} generated after $(k+1)$ steps of the conjugate gradient method satisfies

$$f(x_{k+1}) \leq \left(\frac{b-a}{b+a}\right)^2 f(x_0),$$

where a and b are the smallest and largest eigenvalues of M. Show also that the vector x_{k+1} generated by the scaled conjugate gradient method with $H = M^{-1}$ minimizes f. [*Hint:* Use the interlocking eigenvalues lemma of Luenberger (1973, p. 202).]

The $(k+1)$-step scaled conjugate gradient method is particularly interesting when Q is of the form (77), k is small relative to n, and systems of equations involving M can be solved easily (see Bertsekas, 1974a).

We also leave the following strengthened version of Proposition 1.19 as an exercise to the reader.

Exercise (Hessian with Clustered Eigenvalues): Assume that Q has all its eigenvalues concentrated at k intervals of the form

$$[z_i - \delta_i, z_i + \delta_i], \qquad i = 1, \ldots, k,$$

where we assume that $\delta_i \geq 0, i = 1, \ldots, k, 0 < z_1 - \delta_1$, and

$$0 < z_1 < z_2 < \cdots < z_k, \qquad z_i + \delta_i \leq z_{i+1} - \delta_{i+1}, \qquad i = 1, \ldots, k-1.$$

Show that the vector x_{k+1} generated after $(k+1)$ steps of the conjugate gradient method satisfies

$$f(x_{k+1}) \leq R f(x_0),$$

where

$$R = \max\left\{\frac{\delta_1^2}{z_1^2}, \frac{\delta_2^2(z_2+\delta_2-z_1)^2}{z_1^2 z_2^2}, \frac{\delta_3^2(z_3+\delta_3-z_1)^2(z_3+\delta_3-z_2)^2}{z_1^2 z_2^2 z_3^2}, \right.$$
$$\left. \ldots, \frac{\delta_k^2(z_k+\delta_k-z_1)^2 \cdots (z_k+\delta_k-z_{k-1})^2}{z_1^2 z_2^2 \cdots z_k^2}\right\}.$$

The Conjugate Gradient Method Applied to Nonquadratic Problems

The conjugate gradient method can be applied to the not necessarily quadratic problem

$$\text{minimize} \quad f(x)$$
$$\text{subject to} \quad x \in R^n.$$

It takes the form

$$x_{k+1} = x_k + \alpha_k d_k, \tag{78}$$

where α_k is obtained by line search

$$f(x_k + \alpha_k d_k) = \min_{\alpha} f(x_k + \alpha d_k), \tag{79}$$

and d_k is generated by

$$d_k = -\nabla f(x_k) + \beta_k d_{k-1}. \tag{80}$$

The two most common ways to compute β_k are

$$\beta_k = \frac{\nabla f(x_k)'\nabla f(x_k)}{\nabla f(x_{k-1})'\nabla f(x_{k-1})} \tag{81}$$

and

$$\beta_k = \frac{\nabla f(x_k)'[\nabla f(x_k) - \nabla f(x_{k-1})]}{\nabla f(x_{k-1})'\nabla f(x_{k-1})}. \tag{82}$$

The use of (81) has been suggested by Fletcher and Reeves (1964) while the use of (82) was proposed by Polak and Ribiere (1969), Poljak (1969a), and Sorenson (1969). The direction d_k generated by (80) will be a direction of descent in either case. To see this, note that if $\nabla f(x_k) \neq 0$, then

$$\nabla f(x_k)'d_k = -|\nabla f(x_k)|^2 + \beta_k \nabla f(x_k)'d_{k-1} = -|\nabla f(x_k)|^2 < 0,$$

since $\nabla f(x_k)'d_{k-1} = 0$ in view of (79). However, while these two formulas, along with several others, are equivalent when the method is applied to a quadratic problem, this is no more true in the general case. Extensive computational experience has established that the use of (82) results in much more efficient computation than the use of (81). A heuristic reason that can be given is that due to nonquadratic terms in the objective function and possibly inaccurate line searches, conjugacy of the generated directions is progressively lost and a situation may be created where the method temporarily "jams" in the sense that the generated direction d_k is nearly orthogonal to the gradient $\nabla f(x_k)$. When this occurs, then $\nabla f(x_{k+1}) \simeq \nabla f(x_k)$. In that case β_{k+1}, generated by (82), will be nearly zero and the next direction d_{k+1}, generated by (80), will be close to $-\nabla f(x_{k+1})$ thereby breaking the jam. This is not the case when (81) is used. A more detailed explanation of this phenomenon is given by Powell (1977).

Regardless of the formula for computing the scalar β_k, one must deal with the loss of conjugacy that results from nonquadratic terms in the objective function. The conjugate gradient method is often employed in problems where the number of variables n is large, and it is not unusual

for the method to start generating nonsensical and inefficient directions of search after a few iterations. For this reason it is important to operate the method in cycles of conjugate direction steps given by (80), with the first step in the cycle being a steepest descent step. Some possible restarting policies are:

(a) Restart with a steepest descent step n iterations after the preceding restart.

(b) Restart with a steepest descent step k iterations after the preceding restart with $k < n$. This is recommended when the problem has special structure so that the resulting method has good convergence rate (compare with Proposition 1.19 and the following discussion).

(c) Restart with a steepest descent step n iterations after the preceding restart or if

$$|\nabla f(x_k)'\nabla f(x_{k-1})| > \gamma |\nabla f(x_{k-1})|^2, \tag{83}$$

where γ is a scalar with $0 < \gamma < 1$, whichever comes first. Relation (83) is a test on loss of conjugacy, for if the generated directions were indeed conjugate then we would have $\nabla f(x_k)'\nabla f(x_{k-1}) = 0$. This procedure was suggested by Powell (1977) who recommended the choice of $\gamma = 0.2$.

Note that in all these restart procedures the steepest descent iteration serves as a spacer step and guarantees global convergence (Proposition 1.16). If the scaled version of the conjugate gradient method is used, then a scaled steepest descent iteration is used to restart a cycle. The scaling matrix may change at the beginning of a cycle but should remain unchanged during the cycle. Another possibility, stemming from a suggestion of Beale (1972), is to use the last direction generated in a cycle as the first direction in the new conjugate direction cycle instead of using steepest descent. We refer to papers by Powell (1977) and Shanno (1978a,b) for a discussion of this possibility.

An important practical issue relates to the line search accuracy that is necessary for efficient computation. An elementary calculation shows that if line search is carried out to the extent that

$$\nabla f(x_k)'d_{k-1} < |\nabla f(x_{k-1})|^2,$$

then d_k, generated by (80) and (81), satisfies $\nabla f(x_k)'d_k < 0$ and is a direction of descent. On the other hand, a much more accurate line search may be necessary in order to keep loss of direction conjugacy and deterioration of rate of convergence within a reasonable level. At the same time, insisting on a very accurate line search can be computationally expensive. Considerable research has been directed towards clarifying these questions, and several implementations of the conjugate gradient method with inexact line search have been proposed by Klessig and Polak (1972), Lenard (1973,

1976), and Powell (1977). Among recent works, Shanno (1978a,b) suggests a rather imprecise line search coupled with a method for computing conjugate gradient directions which views each iteration as a memoryless quasi-Newton step. This method appears relatively insensitive to line search errors and yields descent directions under essentially no restriction on line search accuracy.

1.3.5 Quasi-Newton Methods

Quasi-Newton methods are descent methods of the form

$$x_{k+1} = x_k + \alpha_k d_k, \tag{84}$$

$$d_k = -D_k \nabla f(x_k), \tag{85}$$

where D_k is a positive definite matrix adjusted during the course of the computation in a way that (84) tends to approximate Newton's method. The stepsize α_k is determined by one of the stepsize rules of Section 1.3.1. The popularity of the most successful of these methods stems from the fact that they tend to exhibit a fast rate of convergence while avoiding the second derivative calculations associated with Newton's method.

There is a large variety of quasi-Newton methods, but we shall restrict ourselves to the so-called *Broyden class of quasi-Newton algorithms* where D_{k+1} is obtained from D_k and the vectors

$$p_k = x_{k+1} - x_k \tag{86}$$

$$q_k = \nabla f(x_{k+1}) - \nabla f(x_k), \tag{87}$$

by means of the equation

$$D_{k+1} = D_k + \frac{p_k p_k'}{p_k' q_k} - \frac{D_k q_k q_k' D_k}{q_k' D_k q_k} + \zeta_k \tau_k v_k v_k'; \tag{88}$$

where

$$v_k = \frac{P_k}{P_k' q_k} - \frac{D_k q_k}{\tau_k} \tag{89}$$

$$\tau_k = q_k' D_k q_k \tag{90}$$

the scalars ζ_k satisfy, for all k,

$$0 \leq \zeta_k \leq 1, \tag{91}$$

and D_0 is an arbitrary positive definite matrix. If $\zeta_k \equiv 0$, one obtains the *Davidon–Fletcher–Powell (DFP) method* (Davidon, 1959; Fletcher and Powell, 1963), which is historically the first quasi-Newton method. If $\zeta_k \equiv 1$, one obtains the *Broyden–Fletcher–Goldfarb–Shanno (BFGS) method*

(Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970) for which there is growing evidence that it is the best general purpose quasi-Newton method currently available.

We first show that under a mild assumption the matrices D_k generated by (88) are positive definite. This is a most important property, since it guarantees that the search direction d_k is a direction of descent.

Proposition 1.20: If D_k is positive definite, $\nabla f(x_{k+1}) \neq 0$, and the stepsize α_k is chosen so that x_{k+1} satisfies

$$\nabla f(x_k)'d_k < \nabla f(x_{k+1})'d_k \tag{92}$$

(or equivalently $p_k'q_k > 0$), then D_{k+1} given by (88) is well defined and is positive definite.

Proof: First note that (92) implies that $q_k \neq 0$ and

$$p_k'q_k = \alpha_k d_k'[\nabla f(x_{k+1}) - \nabla f(x_k)] > 0. \tag{93}$$

Thus all denominator terms in (88), (89), and (90) are nonzero, and D_{k+1} is well defined.

Now for any $z \neq 0$, we have

$$z'D_{k+1}z = z'D_k z + \frac{(z'p_k)^2}{p_k'q_k} - \frac{(q_k'D_k z)^2}{q_k'D_k q_k} + \zeta_k \tau_k (v_k'z)^2. \tag{94}$$

Define $a = D_k^{1/2}z$, $b = D_k^{1/2}q_k$, and write (94) as

$$z'D_{k+1}z = \frac{|a|^2|b|^2 - (a'b)^2}{|b|^2} + \frac{(z'p_k)^2}{p_k'q_k} + \zeta_k \tau_k (v_k'z)^2. \tag{95}$$

From (90), (91), (93), and the Cauchy–Schwarz inequality we have that all the terms on the right-hand side of (95) are nonnegative. In order that $z'D_{k+1}z > 0$, it will suffice to show that we cannot have simultaneously

$$|a|^2|b|^2 = (a'b)^2 \qquad \text{and} \qquad z'p_k = 0.$$

Indeed if $|a|^2|b|^2 = (a'b)^2$, we must have $a = \lambda b$ for some $\lambda \neq 0$ or $z = \lambda q_k$, so if $z'p_k = 0$, we must have $q_k'p_k = 0$, which is impossible by (93). Q.E.D.

Note that if D_k is positive definite, we have $\nabla f(x_k)'d_k < 0$, so in order to satisfy condition (92), it is sufficient to carry out the line search to a point where

$$|\nabla f(x_{k+1})'d_k| < |\nabla f(x_k)'d_k|.$$

If α_k is determined by the line minimization rule, then $\nabla f(x_{k+1})'d_k = 0$ and (92) is certainly satisfied.

A most interesting property of the Broyden class of algorithms is that when applied to the positive definite quadratic function

$$f(x) = \tfrac{1}{2}x'Qx,$$

with the stepsize α_k determined by line minimization, they generate a Q-conjugate direction sequence, while simultaneously constructing the inverse Hessian Q^{-1} after n iterations. This is the subject of the next proposition.

Proposition 1.21: Let $\{x_k\}$ and $\{d_k\}$ be sequences generated by the algorithm (84)–(90) applied to minimization of the positive definite quadratic function $f(x) = \frac{1}{2}x'Qx$ with α_k chosen by

$$f(x_k + \alpha_k d_k) = \min_{\alpha} f(x_k + \alpha d_k). \tag{96}$$

Assume none of the vectors $x_0, \ldots, x_{n-1}$ is optimal. Then

(a) The vectors $d_0, \ldots, d_{n-1}$ are mutually Q-conjugate.
(b) There holds

$$D_n = Q^{-1}.$$

Proof: It will be sufficient to show that for all k

$$d_i'Qd_j = 0, \qquad 0 \le i < j \le k, \tag{97}$$

$$D_{k+1}q_i = D_{k+1}Qp_i = p_i, \qquad 0 \le i \le k. \tag{98}$$

Equation (97) proves (a). Equation (98) proves (b), since for $k = n - 1$ it shows that $p_0, \ldots, p_{n-1}$ are eigenvectors of D_nQ corresponding to unity eigenvalue. Since $p_i = \alpha_i d_i$ and $d_0, \ldots, d_{n-1}$ are Q-conjugate, it follows that the eigenvectors $p_0, \ldots, p_{n-1}$ are linearly independent and therefore D_nQ equals the identity.

We first verify that for all k

$$D_{k+1}q_k = D_{k+1}Qp_k = p_k. \tag{99}$$

From (88), we have

$$D_{k+1}q_k = D_kq_k + \frac{p_kp_k'q_k}{p_k'q_k} - \frac{D_kq_kq_k'D_kq_k}{q_k'D_kq_k} + \zeta_k\tau_kv_kv_k'q_k = p_k + \zeta_k\tau_kv_kv_k'q_k.$$

An elementary calculation shows that $v_k'q_k = 0$, and (99) follows.

We now show (97) and (98) simultaneously by induction. For $k = 0$ there is nothing to show for (97), while (98) holds in view of (99). Assuming that (97) and (98) hold for k, we prove them for $k + 1$. We have, for $i < k$,

$$\nabla f(x_{k+1}) = \nabla f(x_{i+1}) + Q(p_{i+1} + \cdots + p_k). \tag{100}$$

Using (96), (97), (100), the fact $p_i = \alpha_i d_i$, and the fact $p_k' \nabla f(x_{k+1}) = 0$, we obtain

$$p_i' \nabla f(x_{k+1}) = p_i' \nabla f(x_{i+1}) = 0, \qquad 0 \leq i < k+1.$$

Hence from (98),

$$p_i' Q D_{k+1} \nabla f(x_{k+1}) = 0, \qquad 0 \leq i < k+1,$$

and since $p_i = \alpha_i d_i$, $d_{k+1} = -D_{k+1} \nabla f(x_{k+1})$, we obtain

$$d_i' Q d_{k+1} = 0, \qquad 0 \leq i < k+1.$$

This proves (97) for $k+1$.

From the induction hypothesis (98) and (97), we have

$$q_{k+1}' D_{k+1} q_i = q_{k+1}' D_{k+1} Q p_i = q_{k+1}' p_i = p_{k+1}' Q p_i = 0, \qquad 0 \leq i \leq k. \tag{101}$$

Using (88), (89), (97), (101), and a straightforward calculation, we have, for $0 \leq i \leq k$,

$$\begin{aligned} D_{k+2} q_i &= D_{k+1} q_i + \frac{p_{k+1} p_{k+1}' q_i}{p_{k+1}' q_{k+1}} - \frac{D_{k+1} q_{k+1} q_{k+1}' D_{k+1} q_i}{q_{k+1}' D_{k+1} q_{k+1}} \\ &\quad + \zeta_{k+1} \tau_{k+1} v_{k+1} v_{k+1}' q_i \\ &= D_{k+1} q_i = p_i. \end{aligned}$$

Taking into account (99), we have a proof of (98) for $k+1$. Q.E.D.

It is also interesting to note that *the sequence* $\{x_k\}$ *in Proposition* 1.21 *is identical to the one that would be generated by the scaled conjugate gradient method with scaling matrix* $H = D_0$; i.e., for $k = 0, 1, \ldots, n-1$, the vector x_{k+1} minimizes f over the linear manifold

$$M_k = \{z \mid z = x_0 + \gamma_0 D_0 \nabla f(x_0) + \cdots + \gamma_k D_0 \nabla f(x_k), \gamma_0, \ldots, \gamma_k \in R\}.$$

This can be proved for the case where $D_0 = I$ by verifying by induction that for all k there exist scalars β_{ij}^k such that

$$D_k = I + \sum_{i=0}^{k} \sum_{j=0}^{k} \beta_{ij}^k \nabla f(x_i) \nabla f(x_j)'.$$

Therefore, for some scalars b_i^k and all k, we have

$$d_k = -D_k \nabla f(x_k) = \sum_{i=0}^{k} b_i^k \nabla f(x_i).$$

Hence, for all i, x_{i+1} lies on the manifold

$$M_i = \{z \mid z = x_0 + \gamma_0 \nabla f(x_0) + \cdots + \gamma_i \nabla f(x_i), \gamma_0, \ldots, \gamma_i \in R\},$$

and since the algorithm is a conjugate direction method the result follows using Proposition 1.18. The proof for the case where $D_0 \neq I$ follows by making a transformation of variables so that in the transformed space the initial matrix is the identity. A consequence of this result is that *any algorithm in Broyden's class employing line minimization generates identical sequences of points for the case of a quadratic objective function. This is also true even for a nonquadratic objective function* (Dixon, 1972a,b) which is a rather surprising result. Thus the choice of the scalar ζ_k makes a difference only if the line minimization is inaccurate.

Computational Aspects of Quasi-Newton Methods

Consider now the case of a nonquadratic problem. Even though the quasi-Newton method (84)–(90) is equivalent to the conjugate gradient method for quadratic problems, it has certain advantages which manifest themselves in the presence of inaccurate line search and nonquadratic terms in the objective function. The first advantage is that when line search is accurate the algorithm (84)–(90) not only tends to generate conjugate directions but also constructs an approximation to the inverse Hessian matrix which tends to be more accurate as the algorithm progresses. As a result, near convergence to a strong local minimum, it tends to approximate Newton's method thereby attaining a fast convergence rate. This fact is suggested by Proposition 1.21 and has also been established analytically by Powell (1971) [for a proof, see also Polak (1971)]. It is significant that this property does not depend on the starting matrix D_0, and as a result it is not usually necessary to periodically restart the method with a steepest descent-type step—something that is essential for the conjugate gradient method. A second advantage over the conjugate gradient method is that quasi-Newton methods are not as sensitive to accuracy in the line search. This has been verified by extensive computational experience and can be substantiated to some extent by analysis (see Broyden *et al.*, 1973). One reason that can be given is that, under essentially no restriction on the line search accuracy, the quasi-Newton method (84)–(90) generates positive definite matrices D_k and hence directions of descent (Proposition 1.20).

In an effort to compare further the conjugate gradient method and quasi-Newton methods, we consider their computational requirements per iteration. The kth iteration of the conjugate gradient method requires computation of the objective function and its gradient (perhaps several times in view of the employment of line search) together with $O(n)$† multiplications to compute the conjugate direction d_k and next point x_{k+1}. A

† In this context $O(n)$ multiplications means that there is an integer M such that the number of multiplications per iteration is bounded by Mn, where n is the dimension of the problem.

quasi-Newton method requires roughly the same amount of computation for function and gradient evaluations together with $O(n^2)$ multiplications to compute the matrix D_k and next point x_{k+1}. If the computation time necessary for a function and gradient evaluation is larger or comparable to $O(n^2)$ multiplications, the quasi-Newton method requires only slightly more computation per iteration than the conjugate gradient method and holds the edge in view of its other advantages mentioned earlier. In problems where a function and gradient evaluation requires computation time much less than $O(n^2)$ multiplications, the conjugate gradient method is preferable. For example in optimal control problems where typically n is very large (over 100 and often over 1000) and a function and gradient evaluation typically requires $O(n)$ multiplications, the conjugate gradient method is preferred. In general, both methods require less computation per iteration than Newton's method which requires a function, gradient, and Hessian evaluation, as well as $O(n^3)$ multiplications at each step. This is counterbalanced by the faster speed of convergence of Newton's method. The case for Newton's method is strengthened if periodic reevaluation of the Hessian is employed since each step that utilizes a previously evaluated (and factored) Hessian requires only $O(n^2)$ multiplications. The same is true if the problem has special structure that can be exploited to compute the Newton direction efficiently. For example in optimal control problems, Newton's method typically requires $O(n)$ multiplications per iteration versus $O(n^2)$ multiplications for quasi-Newton methods.

Finally, we note that multiplying the initial matrix D_0 by a positive scaling factor can have a significant beneficial effect on the behavior of the algorithm. A popular choice is to compute

$$\tilde{D}_0 = (p_0' q_0 / q_0' D_0 q_0) D_0 \tag{102}$$

once the vector x_1 (and hence also p_0 and q_0) has been obtained, and use $\tilde{D}_0$ in place of D_0 in computing D_1. The rationale for this is explained in Luenberger (1973). Among other things it can be shown that if the initial scaling (102) is used, then the condition number M_k/m_k, where

$$M_k = \text{max eigenvalue of } (D_k^{1/2} Q D_k^{1/2}),$$

$$m_k = \text{min eigenvalue of } (D_k^{1/2} Q D_k^{1/2}),$$

is not increased (and is usually decreased) at each iteration (compare with the discussion on rate of convergence in Section 1.3.1). Sometimes it is beneficial to scale D_k even after the first iteration by the factor $p_k' q_k / q_k' D_k q_k$ and this has given rise to the class of self-scaling quasi-Newton algorithms due to Oren and Luenberger [see Oren and Luenberger (1974), Oren (1973, 1974), Oren and Spedicato (1976)].

1.3.6 Methods Not Requiring Evaluation of Derivatives

All the gradient methods examined so far in Section 1.3 require calculation of at least the gradient $\nabla f(x_k)$ and possibly the Hessian matrix $\nabla^2 f(x_k)$ at each generated point x_k. In many problems, these derivatives are either not available in explicit form or else are given by very complicated expressions and hence their evaluation requires excessive computation time. In such cases, it is possible to use the same algorithms as earlier with all unavailable derivatives approximated by finite differences. Thus, second derivatives may be approximated by the *forward difference formula*

$$\frac{\partial^2 f(x_k)}{\partial x^i \partial x^j} \sim \frac{1}{h}\left[\frac{\partial f(x_k + he_j)}{\partial x^i} - \frac{\partial f(x_k)}{\partial x^i}\right] \tag{103}$$

or the *central difference formula*

$$\frac{\partial^2 f(x_k)}{\partial x^i \partial x^j} \sim \frac{1}{2h}\left[\frac{\partial f(x_k + he_j)}{\partial x^i} - \frac{\partial f(x_k - he_j)}{\partial x^i}\right]. \tag{104}$$

In these relations, h is a small positive scalar and e_j is the jth unit vector (jth column of the identity matrix). Similarly first derivatives may be approximated by

$$\partial f(x_k)/\partial x^i \sim (1/h)[f(x_k + he_i) - f(x_k)] \tag{105}$$

or by

$$\partial f(x_k)/\partial x^i \sim (1/2h)[f(x_k + he_i) - f(x_k - he_i)]. \tag{106}$$

The central difference formula has the disadvantage that it requires twice as much computation as the forward difference formula. However, it is much more accurate. By forming the corresponding Taylor series expansions, it may be seen that the absolute value of the error between the approximation and the actual derivatives is $O(h)$ for the forward difference formula while it is $O(h^2)$ for the central difference formula. In some cases the same value of h can be used for all partial derivatives, but in other cases, particularly when the problem is poorly scaled, it is essential to use a different value of h for each partial derivative.

From the point of view of reducing the approximation error (or truncation error), it is advantageous to choose the finite difference interval h as small as possible. Unfortunately there is a limit to the amount that h can be reduced due to the significant cancellation error, which occurs when quantities of similar magnitude are subtracted by the computer. Cancellation error is particularly evident in the approximate formulas (105) and (106) near a critical point where ∇f is nearly zero.

Practical experience suggests that a good policy is to keep the scalar h for each derivative at a *fixed* value which balances the truncation error against the cancellation error. When second derivatives are approximated by finite differences of first derivatives in discretized versions of Newton's method, practical experience suggests that extreme accuracy is not very important in terms of speed of convergence. For this reason, exclusive use of the forward difference formula (103) is advisable in most cases. By contrast, when first derivatives are approximated by finite differences of function values, the approximation can become poor near a critical point and can vitally affect the convergence characteristics of the algorithm if the forward difference formula (105) is used exclusively. A good practical rule is to use the forward difference formula (105) until the absolute value of the corresponding approximate derivative becomes less than a certain tolerance; i.e.,

$$|(1/h)[f(x_k + he_i) - f(x_k)]| \leq \varepsilon,$$

where $\varepsilon > 0$ is some small prespecified scalar. At that point a switch to the central difference formula is made; i.e., the formula (106) is used whenever the inequality above is satisfied. This has been suggested by Gill and Murray (1972). An extensive discussion of implementation of gradient methods based on finite difference approximations can be found in Gill *et al.* (1981).

There are several other algorithms for minimizing differentiable functions without the explicit use of derivatives, the most interesting of which, at least from the theoretical point of view, are coordinate descent methods. For a discussion of these and other nonderivative methods we refer the reader to Avriel (1976), Brent (1972), Luenberger (1973), Polak (1971), Powell (1964, 1973), Sargent and Sebastian (1973), and Zangwill (1967a, 1969).

1.4 Constrained Minimization

We consider the problem

$$\text{(CP)} \qquad \begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in X, \end{aligned}$$

where $f: R^n \to R$ is a given function and X is a given subset of R^n. We say that a vector $x^* \in X$ is a *local minimum* for (CP) if there exists an $\varepsilon > 0$ such that

$$f(x^*) \leq f(x) \qquad \forall\, x \in S(x^*; \varepsilon), \quad x \in X.$$

It is a *strict local minimum* if there exists an $\varepsilon > 0$ such that

$$f(x^*) < f(x) \qquad \forall\, x \in S(x^*; \varepsilon), \quad x \in X, \quad x \neq x^*.$$

It is a *global minimum* if

$$f(x^*) \leq f(x) \qquad \forall\, x \in X.$$

We have the following optimality conditions for the case where X is a convex set. Proofs may be found in the sources given at the end of the chapter.

Proposition 1.22: Assume that X is a convex set and for some $\varepsilon > 0$ and $x^* \in X$, $f \in C^1$ over $S(x^*; \varepsilon)$.

(a) If x^* is a local minimum for (CP), then

$$\nabla f(x^*)'(x - x^*) \geq 0 \qquad \forall\, x \in X. \tag{1}$$

(b) If f is in addition convex over X and (1) holds, then x^* is a global minimum for (CP).

We shall be mostly interested in optimality conditions for problems where the constraint set X is described by equality and inequality constraints.

Equality Constrained Problems

We consider first the following equality constrained problem

(ECP) minimize $f(x)$

subject to $h(x) = 0,$

where $f: R^n \to R$ and $h: R^n \to R^m$ are given functions and $m \leq n$. The components of h are denoted $h_1, \ldots, h_m$.

Definition: Let x^* be a vector such that $h(x^*) = 0$ and, for some $\varepsilon > 0$, $h \in C^1$ on $S(x^*; \varepsilon)$. We say that x^* is a *regular point* if the gradients $\nabla h_1(x^*), \ldots, \nabla h_m(x^*)$ are linearly independent.

Consider the *Lagrangian* function $L: R^{n+m} \to R$ defined by

$$L(x, \lambda) = f(x) + \lambda' h(x).$$

We have the following classical results (see, e.g., Luenberger, 1973).

Proposition 1.23: Let x^* be a local minimum for (ECP), and assume that, for some $\varepsilon > 0$, $f \in C^1$, $h \in C^1$ on $S(x^*; \varepsilon)$, and x^* is a regular point. Then there exists a unique vector $\lambda^* \in R^m$ such that

$$\nabla_x L(x^*, \lambda^*) = 0. \tag{2}$$

If in addition $f \in C^2$ and $h \in C^2$ on $S(x^*; \varepsilon)$ then

$$z' \nabla^2_{xx} L(x^*, \lambda^*) z \geq 0 \qquad \forall\, z \in R^n \quad \text{with} \quad \nabla h(x^*)' z = 0. \tag{3}$$

Proposition 1.24: Let x^* be such that $h(x^*) = 0$ and, for some $\varepsilon > 0$, $f \in C^2$ and $h \in C^2$ on $S(x^*; \varepsilon)$. Assume that there exists a vector $\lambda^* \in R^m$ such that

$$\nabla_x L(x^*, \lambda^*) = 0 \tag{4}$$

and

$$z' \nabla^2_{xx} L(x^*, \lambda^*) z > 0 \qquad \forall\, z \neq 0 \quad \text{with} \quad \nabla h(x^*)' z = 0. \tag{5}$$

Then x^* is a strict local minimum for (ECP).

It is instructive to provide a proof of Proposition 1.24 that utilizes concepts that will be of interest later in the analysis of multiplier methods. We have the following lemma:

Lemma 1.25: Let P be a symmetric $n \times n$ matrix and Q a positive semidefinite symmetric $n \times n$ matrix. Assume that $x'Px > 0$ for all $x \neq 0$ satisfying $x'Qx = 0$. Then there exists a scalar c such that

$$P + cQ > 0.$$

Proof: Assume the contrary. Then for every integer k, there exists a vector x_k with $|x_k| = 1$ such that

$$x_k' P x_k + k x_k' Q x_k \leq 0. \tag{6}$$

The sequence $\{x_k\}$ has a subsequence $\{x_k\}_K$ converging to a vector $\bar{x}$ with $|\bar{x}| = 1$. Taking the limit superior in (6), we obtain

$$\bar{x}' P \bar{x} + \limsup_{\substack{k \to \infty \\ k \in K}} (k x_k' Q x_k) \leq 0. \tag{7}$$

Since $x_k' Q x_k \geq 0$, (7) implies that $\{x_k' Q x_k\}_K$ converges to zero and hence $\bar{x}' Q \bar{x} = 0$. From the hypothesis it then follows that $\bar{x}' P \bar{x} > 0$ and this contradicts (7). Q.E.D.

Consider now a vector x^* satisfying the sufficiency assumptions of Proposition 1.24. By Lemma 1.25 it follows that there exists a scalar $\bar{c}$ such that

$$\nabla^2_{xx} L(x^*, \lambda^*) + \bar{c} \nabla h(x^*) \nabla h(x^*)' > 0. \tag{8}$$

Let us introduce the so-called, *augmented Langrangian function*, $L_c: R^{n+m+1} \to R$ defined by

$$L_c(x, \lambda) = f(x) + \lambda' h(x) + \tfrac{1}{2} c |h(x)|^2. \tag{9}$$

We have, by a straightforward calculation,

$$\nabla_x L_c(x, \lambda) = \nabla f(x) + \nabla h(x)[\lambda + ch(x)], \tag{10}$$

$$\nabla^2_{xx} L_c(x, \lambda) = \nabla^2 f(x) + \sum_{i=1}^{m} [\lambda_i + ch_i(x)]\nabla^2 h_i(x) + c\nabla h(x)\nabla h(x)'. \tag{11}$$

Therefore, using also (8), we have, for all $c \geq \bar{c}$,

$$\nabla_x L_c(x^*, \lambda^*) = \nabla_x L(x^*, \lambda^*) = 0, \tag{12}$$

$$\nabla^2_{xx} L_c(x^*, \lambda^*) = \nabla^2_{xx} L(x^*, \lambda^*) + c\nabla h(x^*)\nabla h(x^*)' > 0. \tag{13}$$

Now by using Proposition 1.4 and the preceding discussion, we obtain the following result:

Proposition 1.26: Under the sufficiency assumptions of Proposition 1.24, there exist scalars $\bar{c}, \gamma > 0$, and $\delta > 0$ such that

$$L_c(x, \lambda^*) \geq L_c(x^*, \lambda^*) + \gamma|x - x^*|^2 \qquad \forall\, x \in S(x^*; \delta), \quad c \geq \bar{c}. \tag{14}$$

Notice that from (9) and (14), we obtain

$$f(x) \geq f(x^*) + \gamma|x - x^*|^2 \qquad \forall\, x \in S(x^*; \varepsilon), \quad h(x) = 0,$$

which implies that x^* is a strict local minimum for (ECP). Thus a proof of Proposition 1.24 has been obtained.

The next proposition yields a valuable sensitivity interpretation of Lagrange multipliers. We shall need the following lemma:

Lemma 1.27: Let x^* be a local minimum for (ECP) which is a regular point and together with its associated Lagrange multiplier vector λ^* satisfies the sufficiency assumptions of Proposition 1.24. Then the $(n + m) \times (n + m)$ matrix

$$J = \begin{bmatrix} \nabla^2_{xx} L(x^*, \lambda^*) & \nabla h(x^*) \\ \nabla h(x^*)' & 0 \end{bmatrix} \tag{15}$$

is nonsingular.

Proof: If J were singular, there would exist $y \in R^n$ and $z \in R^m$ not both zero such that (y, z) is in the nullspace of J or equivalently

$$\nabla^2_{xx} L(x^*, \lambda^*)y + \nabla h(x^*)z = 0, \tag{16}$$

$$\nabla h(x^*)'y = 0. \tag{17}$$

Premultiplying (16) by y' and using (17), we obtain

$$y'\nabla^2_{xx} L(x^*, \lambda^*)y = 0.$$

Hence $y = 0$, for otherwise the sufficiency assumption is violated. It follows that $\nabla h(x^*)z = 0$, which in view of the fact that $\nabla h(x^*)$ has rank m implies $z = 0$. This contradicts the fact that y and z cannot be both zero. Q.E.D.

Proposition 1.28: Let the assumptions of Lemma 1.27 hold. Then there exists a scalar $\delta > 0$ and continuously differentiable functions $x(\cdot)\colon S(0;\delta) \to R^n$, $\lambda(\cdot)\colon S(0;\delta) \to R^m$ such that $x(0) = x^*$, $\lambda(0) = \lambda^*$, and for all $u \in S(0;\delta)$, $\{x(u), \lambda(u)\}$ are a local minimum–Lagrange multiplier pair for the problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad h(x) = u. \end{aligned} \tag{18}$$

Furthermore,

$$\nabla_u f[x(u)] = -\lambda(u) \qquad \forall\, u \in S(0;\delta).$$

Proof: Consider the system of equations in (x, λ, u):

$$\nabla f(x) + \nabla h(x)\lambda = 0, \qquad h(x) - u = 0.$$

It has the solution $(x^*, \lambda^*, 0)$. Furthermore the Jacobian of the system with respect to (x, λ) at this solution is the invertible matrix J of (15). Hence by the implicit function theorem (Section 1.2), there exists a $\delta > 0$ and functions $x(\cdot) \in C^1$, $\lambda(\cdot) \in C^1$ on $S(0;\delta)$ such that

$$\nabla f[x(u)] + \nabla h[x(u)]\lambda(u) = 0, \qquad h[x(u)] = u \qquad \forall\, u \in S(0;\delta). \tag{19}$$

For u sufficiently close to $u = 0$, the vectors $x(u)$, $\lambda(u)$ satisfy the sufficiency conditions for problem (18) in view of the fact that they satisfy them by assumption for $u = 0$. Hence δ can be chosen so that $\{x(u), \lambda(u)\}$ are a local minimum–Lagrange multiplier pair for problem (18).

Now from (19), we have

$$\nabla_u x(u)\nabla f[x(u)] + \nabla_u x(u)\nabla h[x(u)]\lambda(u) = 0$$

or

$$\nabla_u f[x(u)] = -\nabla_u x(u)\nabla h[x(u)]\lambda(u). \tag{20}$$

By differentiating the relation $h[x(u)] = u$, we obtain

$$I = \nabla_u h[x(u)] = \nabla_u x(u)\nabla h[x(u)]. \tag{21}$$

Combining (20) and (21), we have

$$\nabla_u f[x(u)] = -\lambda(u),$$

which was to be proved. Q.E.D.

Inequality Constraints

Consider now the case of a problem involving both equality and inequality constraints

$$\text{(NLP)} \qquad \begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad h(x) = 0, \qquad g(x) \leq 0, \end{aligned}$$

where $f\colon R^n \to R$, $h\colon R^n \to R^m$, $g\colon R^n \to R^r$ are given functions and $m \leq n$. The components of g are denoted by $g_1, \ldots, g_r$. We first generalize the definition of a regular point. For any vector x satisfying $g(x) \leq 0$, we denote

$$A(x) = \{j \mid g_j(x) = 0, j = 1, \ldots, r\}. \tag{22}$$

Definition: Let x^* be a vector such that $h(x^*) = 0$, $g(x^*) \leq 0$ and, for some $\varepsilon > 0$, $h \in C^1$ and $g \in C^1$ on $S(x^*; \varepsilon)$. We say that x^* is a *regular point* if the gradients $\nabla h_1(x^*), \ldots, \nabla h_m(x^*)$ and $\nabla g_j(x^*)$, $j \in A(x^*)$, are linearly independent.

Define the Lagrangian function $L\colon R^{n+m+r} \to R$ for (NLP) by

$$L(x, \lambda, \mu) = f(x) + \lambda' h(x) + \mu' g(x).$$

We have the following optimality conditions paralleling those for equality constrained problems (see, e.g., Luenberger, 1973).

Proposition 1.29: Let x^* be a local minimum for (NLP) and assume that, for some $\varepsilon > 0$, $f \in C^1$, $h \in C^1$, $g \in C^1$ on $S(x^*; \varepsilon)$, and x^* is a regular point. Then there exist unique vectors $\lambda^* \in R^m$, $\mu^* \in R^r$ such that

$$\nabla_x L(x^*, \lambda^*, \mu^*) = 0, \tag{23}$$

$$\mu_j^* \geq 0, \qquad \mu_j^* g_j(x^*) = 0 \qquad \forall j = 1, \ldots, r. \tag{24}$$

If in addition $f \in C^2$, $h \in C^2$, and $g \in C^2$ on $S(x^*; \varepsilon)$, then for all $z \in R^n$ satisfying $\nabla h(x^*)' z = 0$ and $\nabla g_j(x^*)' z = 0, j \in A(x^*)$, we have

$$z' \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) z \geq 0. \tag{25}$$

Proposition 1.30: Let x^* be such that $h(x^*) = 0$, $g(x^*) \leq 0$, and, for some $\varepsilon > 0$, $f \in C^2$, $h \in C^2$, and $g \in C^2$ on $S(x^*; \varepsilon)$. Assume that there exist vectors $\lambda^* \in R^m$, $\mu^* \in R^r$ such that

$$\nabla_x L(x^*, \lambda^*, \mu^*) = 0, \tag{26}$$

$$\mu_j^* \geq 0, \qquad \mu_j^* g_j(x^*) = 0 \qquad \forall j = 1, \ldots, r, \tag{27}$$

and for every $z \neq 0$ satisfying $\nabla h(x^*)'z = 0$, $\nabla g_j(x^*)'z \leq 0$, for all $j \in A(x^*)$, and $\nabla g_j(x^*)'z = 0$, for all $j \in A(x^*)$ with $\mu_j^* > 0$, we have

$$z'\nabla_{xx}^2 L(x^*, \lambda^*, \mu^*)z > 0. \tag{28}$$

Then x^* is a strict local minimum for (NLP).

Optimality Conditions via Conversion to the Equality Constrained Case

Some of the results for inequality constraints may also be proved by using the results for equality constraints *provided we assume that* $f, h_i, g_j \in C^2$. In this approach, we convert the inequality constrained problem (NLP) into a problem which involves exclusively equality constraints and then use the results for (ECP) to obtain necessary conditions, sufficiency conditions, and a sensitivity result for (NLP).

Consider the equality constrained problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad h_1(x) = 0, \ldots, h_m(x) = 0, \\ &\qquad\qquad g_1(x) + z_1^2 = 0, \ldots, g_r(x) + z_r^2 = 0, \end{aligned} \tag{29}$$

where we have introduced additional variables $z_1, \ldots, z_r$. It is clear that (NLP) and problem (29) are equivalent in the sense that x^* is a local minimum for problem (NLP) if and only if $(x^*, [-g_1(x^*)]^{1/2}, \ldots, [-g_r(x^*)]^{1/2})$ is a local minimum for (29). By introducing the vector $z = (z_1, \ldots, z_r)$ and the functions

$$\begin{aligned} \bar{f}(x, z) &= f(x), \\ \bar{h}_i(x, z) &= h_i(x), \qquad i = 1, \ldots, m, \\ \bar{g}_j(x, z) &= g_j(x) + z_j^2, \qquad j = 1, \ldots, r, \end{aligned}$$

problem (29) may be written as

$$\begin{aligned} &\text{minimize} \quad \bar{f}(x, z) \\ &\text{subject to} \quad \bar{h}_i(x, z) = 0, \quad \bar{g}_j(x, z) = 0, \quad i = 1, \ldots, m, \quad j = 1, \ldots, r. \end{aligned} \tag{30}$$

Let x^* be a local minimum for our original problem (NLP) as well as a regular point. Then (x^*, z^*), where $z^* = (z_1^*, \ldots, z_r^*)$, $z_j^* = [-g_j(x^*)]^{1/2}$,

is a local minimum for problem (30). In addition (x^*, z^*) is a regular point since the gradients

$$\nabla \bar{h}_i(x^*, z^*) = \begin{bmatrix} \nabla h_i(x^*) \\ 0 \end{bmatrix}, \qquad i = 1, \ldots, m,$$

$$\nabla \bar{g}_j(x^*, z^*) = \begin{bmatrix} \nabla g_j(x^*) \\ 0 \\ \vdots \\ 0 \\ 2z_j^* \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad j = 1, \ldots, r,$$

can be easily verified to be linearly independent when x^* is a regular point. By the necessary conditions for equality constraints (Proposition 1.23), there exist Lagrange multipliers $\lambda_1^*, \ldots, \lambda_m^*, \mu_1^*, \ldots, \mu_r^*$ such that

$$\nabla \bar{f}(x^*, z^*) + \sum_{i=1}^{m} \lambda_i^* \nabla \bar{h}_i(x^*, z^*) + \sum_{j=1}^{r} \mu_j^* \nabla \bar{g}_j(x^*, z^*) = 0.$$

In view of the form of the gradients of $\bar{f}$, $\bar{h}_i$, and $\bar{g}_j$, the condition above is equivalent to

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^{r} \mu_j^* \nabla g_j(x^*) = 0, \tag{31a}$$

$$2\mu_j^* [-g_j(x^*)]^{1/2} = 0, \qquad j = 1, \ldots, r. \tag{31b}$$

The last equation implies $\mu_j^* = 0$ for all $j \notin A(x^*)$ and may also be written as

$$\mu_j^* g_j(x^*) = 0, \qquad j = 1, \ldots, r. \tag{32}$$

The second-order necessary condition for problem (30) is applicable, in view of our assumption $f, h_i, g_j \in C^2$ which in turn implies $\bar{f}, \bar{h}_i, \bar{g}_j \in C^2$. It yields

$$[y', v'] \left[\begin{array}{c|ccc} \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) & & 0 & \\ \hline & 2\mu_1^* & & 0 \\ 0 & & \ddots & \\ & 0 & & 2\mu_r^* \end{array}\right] \begin{bmatrix} y \\ v \end{bmatrix} \geq 0 \tag{33}$$

for all $y \in R^n$, $v = (v_1, \ldots, v_r) \in R^r$ satisfying

$$\nabla h(x^*)'y = 0, \qquad \nabla g_j(x^*)'y + 2z_j^* v_j = 0, \qquad j = 1, \ldots, r. \tag{34}$$

By setting $v_j = 0$ for $j \in A(x^*)$ and taking into account the fact $\mu_j^* = 0$ for $j \notin A(x^*)$ [compare with (32)] we obtain, from (33) and (34),

$$y'\nabla_{xx}^2 L(x^*, \lambda^*, \mu^*)y \geq 0, \tag{35}$$

$$\forall\, y, \text{ with } \nabla h(x^*)'y = 0, \quad \nabla g_j(x^*)'y = 0, \quad j \in A(x^*).$$

For every j with $z_j^* = 0$, we may choose $y = 0$, $v_j \neq 0$, and $v_k = 0$, for $k \neq j$, in (33) to obtain

$$\mu_j^* \geq 0. \tag{36}$$

Relations (31), (32), (35), and (36) represent all the necessary conditions of Proposition 1.29. Thus we have obtained a proof of Proposition 1.29 (under the assumption f, h_i, $g_j \in C^2$) based on the transformation of the inequality constrained problem (NLP) to the equality constrained problem (29).

The transformation described above may also be used to derive a set of sufficiency conditions for (NLP) which are somewhat weaker than those of Proposition 1.30.

Proposition 1.31: Let x^* be such that $h(x^*) = 0$, $g(x^*) \leq 0$, and, for some $\varepsilon > 0$, $f \in C^2$, $h \in C^2$, and $g \in C^2$ on $S(x^*; \varepsilon)$. Assume that there exist vectors $\lambda^* \in R^m$, $\mu^* \in R^r$ satisfying

$$\nabla_x L(x^*, \lambda^*, \mu^*) = 0,$$

$$\mu_j^* \geq 0, \qquad \mu_j^* g_j(x^*) = 0, \qquad j = 1, \ldots, r,$$

as well as the *strict complementarity condition*

$$\mu_j^* > 0 \qquad \text{if} \quad j \in A(x^*).$$

Assume further that for all $y \neq 0$ satisfying $\nabla h(x^*)'y = 0$ and $\nabla g_j(x^*)'y = 0$, for all $j \in A(x^*)$, we have

$$y'\nabla_{xx}^2 L(x^*, \lambda^*, \mu^*)y > 0.$$

Then x^* is a strict local minimum for (NLP).

Proof: From (31), (33), and (34), we see that our assumptions imply that the sufficiency conditions of Proposition 1.24 are satisfied for (x^*, z^*) and λ^*, μ^*, where $z^* = ([-g_1(x^*)]^{1/2}, \ldots, [-g_r(x^*)]^{1/2})$ for problem (29). Hence (x^*, z^*) is a strict local minimum for problem (29) and it follows that x^* is a strict local minimum of f subject to $h(x) = 0$, and $g(x) \leq 0$. Q.E.D.

We formalize some of the arguments in the preceding discussion in the following proposition.

Proposition 1.32: If the sufficiency conditions for (NLP) of Proposition 1.31 hold, then the sufficiency conditions of Proposition 1.24 are satisfied for problem (29). If in addition x^* is a regular point for (NLP), then (x^*, z^*), where $z^* = ([-g_1(x^*)]^{1/2}, \ldots, [-g_r(x^*)]^{1/2})$, is a regular point for problem (29).

Linear Constraints

The preceding necessary conditions rely on a regularity assumption on the local minimum x^* to assert the existence of a unique Lagrange multiplier vector. When x^* is not regular, there are two possibilities. Either there does not exist a Lagrange multiplier vector or there exists an infinity of such vectors. There are a number of assumptions other than regularity that guarantee the existence of a Lagrange multiplier vector. A very useful one is linearity of the constraint functions as in the following proposition.

Proposition 1.33: Let x^* be a local minimum for the problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad a_j'x - b_j \leq 0, \qquad j = 1, \ldots, r, \end{aligned}$$

where $f: R^n \to R$, $b \in R^r$, and $a_j \in R^n$, $j = 1, \ldots, r$. Assume that, for some $\varepsilon > 0$, $f \in C^1$ on $S(x^*; \varepsilon)$. Then there exists a vector $\mu^* = (\mu_1^*, \ldots, \mu_r^*)$ such that

$$\begin{aligned} &\nabla f(x^*) + \sum_{j=1}^{r} \mu_j^* a_j = 0, \\ &\mu_j^* \geq 0, \qquad \mu_j^*(a_j'x^* - b_j) = 0, \qquad j = 1, \ldots, r. \end{aligned}$$

Sufficiency Conditions under Convexity Assumptions

Consider the convex programming problem

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad g(x) \leq 0, \end{aligned} \tag{37}$$

where we assume that the functions f and $g_1, \ldots, g_r$ are convex and differentiable over R^n. Then every local minimum is global, and the necessary optimality conditions of Proposition 1.29 are also sufficient as stated in the following proposition.

Proposition 1.34: Assume that f and $g_1, \ldots, g_r$ are convex and continuously differentiable functions on R^n. Let $x^* \in R^n$ and $\mu^* \in R^r$ satisfy

$$\nabla f(x^*) + \nabla g(x^*)\mu^* = 0,$$

$$g(x^*) \leq 0, \qquad \mu_j^* \geq 0, \qquad \mu_j^* g_j(x^*) = 0, \qquad j = 1, \ldots, r.$$

Then x^* is a global minimum of problem (37).

1.5 Algorithms for Minimization Subject to Simple Constraints

There is a large number of algorithms of the feasible direction type for minimization of differentiable functions subject to linear constraints. A survey of some of the most popular ones may be found in the volume edited by Gill and Murray (1974), and computational results may be found in the paper by Lenard (1979). In this section, we shall focus on a new class of methods that is well suited for problems with simple inequality constraints such as those that might arise in methods of multipliers and differentiable exact penalty methods, where the simple constraints are not eliminated by means of a penalty but rather are treated directly (cf. Sections 2.4 and 4.3). We shall restrict ourselves exclusively to problems involving lower and/or upper bounds on the variables, but there are extensions of the class of algorithms presented that handle problems with general linear constraints (see Bertsekas, 1980c).

Consider the problem

$$\text{(SCP)} \qquad \begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \geq 0, \end{aligned}$$

where f: $R^n \to R$ is a continuously differentiable function. By applying Proposition 1.22, we obtain the following necessary conditions for optimality of a vector $x^* \geq 0$.

$$\text{(1a)} \qquad \partial f(x^*)/\partial x^i = 0 \qquad \text{if} \quad x^{*i} > 0, \quad i = 1, \ldots, n,$$

$$\text{(1b)} \qquad \partial f(x^*)/\partial x^i \geq 0 \qquad \text{if} \quad x^{*i} = 0, \quad i = 1, \ldots, n.$$

An equivalent way of writing these conditions is

$$\text{(2)} \qquad x^* = [x^* - \alpha \nabla f(x^*)]^+,$$

where α is any positive scalar and $[\cdot]^+$ denotes projection on the positive orthant; i.e., for every $z = (z^1, \ldots, z^n)$,

$$\text{(3)} \qquad [z]^+ = \begin{bmatrix} \max\{0, z^1\} \\ \vdots \\ \max\{0, z^n\} \end{bmatrix}.$$

If a vector $x^* \geq 0$ satisfies (1), we say that it is a *critical point* with respect to (SCP).

Equation (2) motivates the following extension of the steepest descent method

$$x_{k+1} = [x_k - \alpha_k \nabla f(x_k)]^+, \qquad k = 0, 1, \ldots, \tag{4}$$

where α_k is a positive scalar stepsize. There are a number of rules for choosing α_k that guarantee that limit points of sequences generated by iteration (4) satisfy the necessary condition (1) (Goldstein, 1964, 1974; Levitin and Poljak, 1965; McCormick, 1969; Bertsekas, 1974c). The rate of convergence of iteration (4) is however at best linear for general problems. We shall provide Newton-like generalizations of iteration (4) which preserve its basic simplicity while being capable of superlinear convergence.

Consider an iteration of the form

$$x_{k+1} = [x_k - \alpha_k D_k \nabla f(x_k)]^+, \qquad k = 0, 1, \ldots, \tag{5}$$

where D_k is a positive definite symmetric matrix and α_k is chosen by search along the arc of points

$$x_k(\alpha) = [x_k - \alpha D_k \nabla f(x_k)]^+, \qquad \alpha \geq 0. \tag{6}$$

It is easy to construct examples (see Fig. 1.2) where an arbitrary choice of the matrix D_k leads to situations where it is impossible to reduce the value

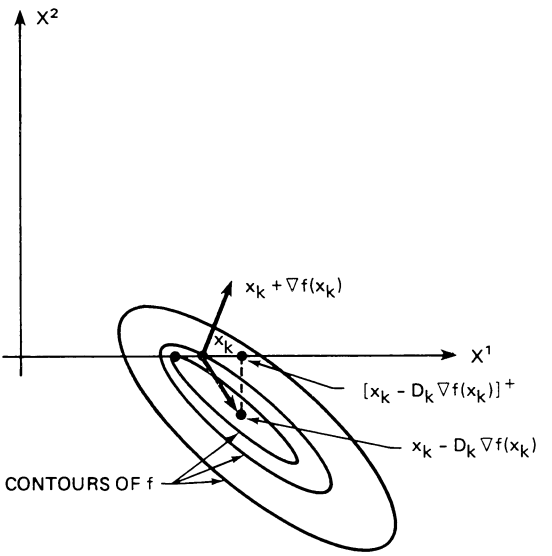

FIG. 1.2

of the objective by suitable choice of the stepsize α (i.e., $f[x_k(\alpha)] \geq f(x_k)$ $\forall\, \alpha \geq 0$). The following proposition identifies a class of matrices D_k for which an objective reduction is possible. Define, for all $x \geq 0$,

$$I^+(x) = \{i \mid x^i = 0, \partial f(x)/\partial x^i > 0\}. \tag{7}$$

We say that a symmetric matrix D with elements d^{ij} is *diagonal with respect to a subset of indices* $I \subset \{1, 2, \ldots, n\}$, if

$$d^{ij} = 0 \qquad \forall\, i \in I, \quad j = 1, 2, \ldots, n, \quad j \neq i. \tag{8}$$

Proposition 1.35: Let $x \geq 0$ and D be a positive definite symmetric matrix which is diagonal with respect to $I^+(x)$, and denote

$$x(\alpha) = [x - \alpha D \nabla f(x)]^+ \qquad \forall\, \alpha \geq 0. \tag{9}$$

(a) The vector x is a critical point with respect to (SCP), if and only if

$$x = x(\alpha) \qquad \forall\, \alpha \geq 0.$$

(b) If x is not a critical point with respect to (SCP), there exists a scalar $\bar{\alpha} > 0$ such that

$$f[x(\alpha)] < f(x) \qquad \forall\, \alpha \in (0, \bar{\alpha}]. \tag{10}$$

Proof: Assume without loss of generality that for some integer r, we have

$$I^+(x) = \{r + 1, \ldots, n\}.$$

Then D has the form

$$D = \begin{bmatrix} \bar{D} & \vdots & & 0 \\ \cdots & \vdots & d^{r+1} & \\ & \vdots & & \ddots \\ 0 & \vdots & 0 & & d^n \end{bmatrix}, \tag{11}$$

where $\bar{D}$ is positive definite and $d^i > 0$, $i = r + 1, \ldots, n$.

Denote

$$p = D \nabla f(x). \tag{12}$$

(a) Assume x is a critical point. Then, using (1), (7),

$$\partial f(x)/\partial x^i = 0 \qquad \forall\, i = 1, \ldots, r$$
$$\partial f(x)/\partial x^i > 0, \qquad x^i = 0 \qquad \forall\, i = r + 1, \ldots, n.$$

These relations and the positivity of d^i, $i = r + 1, \ldots, n$, imply that

$$p^i = 0 \qquad \forall\, i = 1, \ldots, r,$$
$$p^i > 0 \qquad \forall\, i = r + 1, \ldots, n.$$

Since $x^i(\alpha) = [x^i - \alpha p^i]^+$ and $x^i = 0$ for $i = r+1, \ldots, n$, it follows that $x^i(\alpha) = x^i$, for all i, and $\alpha \geq 0$.

Conversely assume that $x = x(\alpha)$ for all $\alpha \geq 0$. Then we must have

$$p^i = 0 \qquad \forall\, i = 1, \ldots, n \quad \text{with} \quad x^i > 0,$$

$$p^i \geq 0 \qquad \forall\, i = 1, \ldots, n \quad \text{with} \quad x^i = 0.$$

Now by definition of $I^+(x)$, we have that if $x^i = 0$ and $i \notin I^+(x)$, then $\partial f(x)/\partial x^i \leq 0$. This together with the relations above imply

$$\sum_{i=1}^{r} p^i \frac{\partial f(x)}{\partial x^i} \leq 0.$$

Since, by (11) and (12),

$$\begin{bmatrix} p_1 \\ \vdots \\ p_r \end{bmatrix} = \bar{D} \begin{bmatrix} \partial f(x)/\partial x^1 \\ \vdots \\ \partial f(x)/\partial x^r \end{bmatrix}$$

and $\bar{D}$ is positive definite, it follows that

$$p^i = \partial f(x)/\partial x^i = 0 \qquad \forall\, i = 1, \ldots, r.$$

Since, for $i = r+1, \ldots, n$, $\partial f(x)/\partial x^i > 0$, and $x^i = 0$, we obtain that x is a critical point.

(b) For $i = r+1, \ldots, n$, we have $\partial f(x)/\partial x^i > 0$, $x^i = 0$, and, from (11) and (12), $p^i > 0$. Since $x^i(\alpha) = [x^i - \alpha p^i]^+$, we obtain

$$x^i = x^i(\alpha) = 0 \qquad \forall\, \alpha \geq 0, \quad i = r+1, \ldots, n. \tag{13}$$

Consider the sets of indices

$$I_1 = \{i \mid x^i > 0 \quad \text{or} \quad x^i = 0 \quad \text{and} \quad p^i < 0, \quad i = 1, \ldots, r\}, \tag{14}$$

$$I_2 = \{i \mid x^i = 0 \quad \text{and} \quad p^i \geq 0, \quad i = 1, \ldots, r\}. \tag{15}$$

Let

$$\alpha_1 = \sup\{\alpha \geq 0 \mid x^i - \alpha p^i \geq 0, i \in I_1\}. \tag{16}$$

Note that, in view of the definition of I_1, α_1 is either positive or $+\infty$. Define the vector $\bar{p}$ with coordinates

$$\bar{p}^i = \begin{cases} p^i & \text{if} \quad i \in I_1, \\ 0 & \text{if} \quad i \in I_2 \quad \text{or} \quad i = r+1, \ldots, n. \end{cases} \tag{17}$$

In view of (13)–(16), we have

$$x(\alpha) = x - \alpha \bar{p} \qquad \forall\, \alpha \in (0, \alpha_1). \tag{18}$$

In view of (15) and the definition of $I^+(x)$, we have

$$\partial f(x)/\partial x^i \leq 0 \qquad \forall\, i \in I_2, \tag{19}$$

and hence

$$\sum_{i \in I_2} \frac{\partial f(x)}{\partial x^i} p^i \leq 0. \tag{20}$$

Now using (17) and (20), we have

$$\nabla f(x)'\bar{p} = \sum_{i \in I_1} \frac{\partial f(x)}{\partial x^i} p^i \geq \sum_{i=1}^{r} \frac{\partial f(x)}{\partial x^i} p^i. \tag{21}$$

Since x is not a critical point, by part (a) and (18), we must have $x \neq x(\alpha)$ for some $\alpha > 0$, and hence also in view of (13), $p^i \neq 0$ for some $i \in \{1, \ldots, r\}$. In view of the positive definiteness of $\bar{D}$ and (11) and (12), it follows that

$$\sum_{i=1}^{r} \frac{\partial f(x)}{\partial x^i} p^i > 0.$$

It follows, from (21), that

$$\nabla f(x)'\bar{p} > 0.$$

Combining this relation with (18) and the fact that $\alpha_1 > 0$, it follows that $\bar{p}$ is a feasible descent direction at x and there exists a scalar $\bar{\alpha} > 0$ for which the desired relation (10) is satisfied. Q.E.D.

Based on Proposition 1.35, we are led to the conclusion that the matrix D_k in the iteration

$$x_{k+1} = [x_k - \alpha_k D_k \nabla f(x_k)]^+$$

should be chosen diagonal with respect to a subset of indices that contains

$$I^+(x_k) = \{i \mid x_k^i = 0, \partial f(x_k)/\partial x^i > 0\}.$$

Unfortunately, the set $I^+(x_k)$ exhibits an undesirable discontinuity at the boundary of the constraint set whereby given a sequence $\{x_k\}$ of interior points that converges to a boundary point $\bar{x}$, all the sets $I^+(x_k)$ may be strictly smaller than the set $I^+(\bar{x})$. This causes difficulties in proving convergence of the algorithm and may have an adverse effect on its rate of convergence. (This phenomenon is quite common in feasible direction algorithms and is referred to as zigzagging or jamming.) For this reason, we shall employ certain enlargements of the sets $I^+(x_k)$ with the aim of bypassing these difficulties.

The algorithm that we describe utilizes a scalar $\varepsilon > 0$ (typically small), a fixed† *diagonal* positive definite matrix M (for example, the identity), and two parameters $\beta \in (0, 1)$ and $\sigma \in (0, \frac{1}{2})$ that will be used in connection with an Armijo-like stepsize rule. An initial vector $x_0 \geq 0$ is chosen and at the kth iteration of the algorithm, we have a vector $x_k \geq 0$. Denote

$$w_k = |x_k - [x_k - M\nabla f(x_k)]^+|, \qquad \varepsilon_k = \min\{\varepsilon, w_k\}.$$

(Actually there are several other possibilities for defining the scalar ε_k as can be seen by examination of the proof of the subsequent proposition. It is also possible to use a separate scalar ε_k^i for each coordinate.)

$(k + 1)$st Iteration of the Algorithm

We select a positive definite symmetric matrix D_k which is diagonal with respect to the set I_k^+ given by

$$I_k^+ = \{i \mid 0 \leq x_k^i \leq \varepsilon_k, \partial f(x_k)/\partial x^i > 0\}. \tag{22}$$

Denote

$$p_k = D_k \nabla f(x_k), \tag{23}$$

$$x_k(\alpha) = [x_k - \alpha p_k]^+ \qquad \forall\, \alpha \geq 0. \tag{24}$$

Then x_{k+1} is given by

$$x_{k+1} = x_k(\alpha_k), \tag{25}$$

where

$$\alpha_k = \beta^{m_k}, \tag{26}$$

and m_k is the first nonnegative integer m such that

$$f(x_k) - f[x_k(\beta^m)] \geq \sigma\left\{\beta^m \sum_{i \notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i + \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i}[x_k^i - x_k^i(\beta^m)]\right\}. \tag{27}$$

The stepsize rule (26) and (27) is quite similar to the Armijo rule of Section 1.3. We have chosen a unity initial stepsize, but any other positive initial stepsize can be incorporated in the matrix D_k, so this choice involves no loss of generality. The results that follow can also be proved if

$$\sum_{i \notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i$$

† Actually the results that follow can also be proved if the fixed matrix M is replaced by a sequence of diagonal positive definite matrices $\{M_k\}$ with diagonal elements that are bounded above and away from zero.

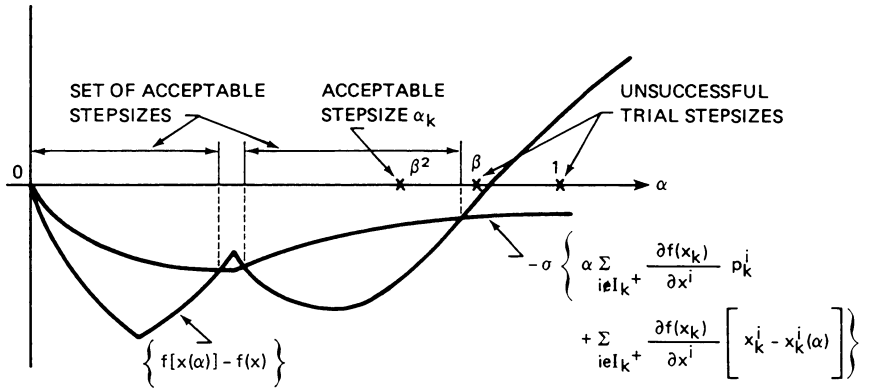

FIG. 1.3 Line search by the Armijo-like rule (26), (27).

in (27) is replaced by $\gamma_k \sum_{i\notin I_k^-}[\partial f(x_k)/\partial x^i]p_k^i$, where $\gamma_k = \min\{1, \bar{\alpha}_k\}$ and $\bar{\alpha}_k = \sup\{\alpha \mid x_k^i - \alpha p_k^i \geq 0 \; \forall\, i \notin I_k^+\}$. Other variations of the stepsize rule are also possible. The process of determining the stepsize α_k is illustrated in Fig. 1.3. When I_k^+ is empty, the right-hand side of (27) becomes $\sigma\beta^m \nabla f(x_k)' p_k$ and is identical to the corresponding expression of the Armijo rule for unconstrained minimization. Note that, for all k, $I_k^+ \supset I^+(x_k)$ so D_k is diagonal with respect to $I^+(x_k)$. It is possible to show that for all $m \geq 0$, the right-hand side of (27) is nonnegative and is positive if and only if x_k is not a critical point. Indeed since D_k is positive definite and diagonal with respect to I_k^+, we have

$$\sum_{i\notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i \geq 0 \qquad \forall\, k = 0, 1, \ldots, \tag{28}$$

while for all $i \in I_k^+$, in view of the fact $\partial f(x^k)/\partial x^i > 0$, we have $p_k^i > 0$, and hence

$$x_k^i - x_k^i(\alpha) \geq 0 \qquad \forall\, \alpha \geq 0, \quad i \in I_k^+, \quad k = 0, 1, \ldots,$$

$$\frac{\partial f(x_k)}{\partial x^i}[x_k^i - x_k^i(\alpha)] \geq 0 \qquad \forall\, \alpha \geq 0, \quad i \in I_k^+, \quad k = 0, 1, \ldots. \tag{29}$$

This shows that the right-hand side of (27) is nonnegative. If x_k is not critical, then it is easily seen [compare also with the proof of Proposition 1.35(b)] that one of the inequalities (28) or (29) is strict for $\alpha > 0$ so the right-hand side of (27) is positive for all $m \geq 0$. A slight modification of the proof of Proposition 1.35(b) also shows that if x_k is not a critical point, then (27) will be satisfied for all m sufficiently large so the stepsize α_k is well defined and can be determined via a finite number of arithmetic operations. If x_k is a critical point then, by Proposition 1.35(a), we have $x_k = x_k(\alpha)$ for all $\alpha \geq 0$.

Furthermore the argument given in the proof of Proposition 1.35(a) shows that

$$\sum_{i \notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i = 0,$$

so both terms in the right-hand side of (27) are zero. Since also $x_k = x_k(\alpha)$ for all $\alpha \geq 0$, it follows that (27) is satisfied for $m = 0$ thereby implying that

$$x_{k+1} = x_k(1) = x_k \quad \text{if} \quad x_k \text{ is critical.}$$

In conclusion the algorithm is well defined, decreases the value of the objective function at each iteration k for which x_k is not a critical point, and essentially terminates if x_k is critical. We proceed to analyze its convergence and rate of convergence properties. To this end, we shall make use of the following two assumptions:

Assumption (A): *The gradient ∇f is Lipschitz continuous on each bounded set of R^n; i.e., given any bounded set $S \subset R^n$ there exists a scalar L (depending on S) such that*

$$|\nabla f(x) - \nabla f(y)| \leq L|x - y| \qquad \forall\, x, y \in S. \tag{30}$$

Assumption (B): *There exist positive scalars λ_1 and λ_2 and nonnegative integers q_1 and q_2, such that*

$$\lambda_1 w_k^{q_1} |z|^2 \leq z' D_k z \leq \lambda_2 w_k^{q_2} |z|^2 \qquad \forall\, z \in R^n, \quad k = 0, 1, \ldots, \tag{31}$$

where

$$w_k = |x_k - [x_k - M\nabla f(x_k)]^+|.$$

Assumption (A) is not essential for the result of Proposition 1.36 that follows but simplifies its proof. It is satisfied for just about every problem likely to appear in practice. For example, it is satisfied when f is twice differentiable, as well as when f is an augmented Lagrangian of the type considered in Chapter 3 for problems involving twice differentiable functions. Assumption (B) is a condition of the type utilized in connection with unconstrained minimization algorithms (compare with the discussion preceding Proposition 1.8). When $q_1 = q_2 = 0$, relation (31) takes the form

$$\lambda_1 |z|^2 \leq z' D_k z \leq \lambda_2 |z|^2 \qquad \forall\, z \in R^n, \quad k = 0, 1, \ldots, \tag{32}$$

and simply says that the eigenvalues of D_k are uniformly bounded above and away from zero.

Proposition 1.36: Under Assumptions (A) and (B) above, every limit point of a sequence $\{x_k\}$ generated by iteration (25) is a critical point with respect to (SCP).

Proof: Assume the contrary; i.e., there exists a subsequence $\{x_k\}_K$ converging to a vector $\bar{x}$ which is not critical. Since $\{f(x_k)\}$ is decreasing and f is continuous, it follows that $\{f(x_k)\}$ converges to $f(\bar{x})$ and therefore

$$[f(x_k) - f(x_{k+1})] \to 0.$$

Since each of the sums in the right-hand side of (27) is nonnegative [compare with (28) and (29)], we must have

$$\alpha_k \sum_{i \notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i \to 0, \tag{33}$$

$$\sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} [x_k^i - x_k^i(\alpha_k)] \to 0. \tag{34}$$

Also since $\bar{x}$ is not critical and M is positive definite and diagonal, we have clearly $|\bar{x} - [\bar{x} - M\nabla f(\bar{x})]^+| \neq 0$, so (31) implies that the eigenvalues of $\{D_k\}_K$ are uniformly bounded above and away from zero. In view of the fact that D_k is diagonal with respect to I_k^+, it follows that there exist positive scalars $\bar{\lambda}_1$ and $\bar{\lambda}_2$ such that, for all $k \in K$ that are sufficiently large,

$$0 < \bar{\lambda}_1 \, \partial f(x_k)/\partial x^i \leq p_k^i \leq \bar{\lambda}_2 \, \partial f(x_k)/\partial x^i \qquad \forall\, i \in I_k^+, \tag{35}$$

$$\bar{\lambda}_1 \sum_{i \notin I_k^+} \left| \frac{\partial f(x_k)}{\partial x^i} \right|^2 \leq \sum_{i \notin I_k^+} p_k^i \frac{\partial f(x_k)}{\partial x^i} \leq \bar{\lambda}_2 \sum_{i \notin I_k^+} \left| \frac{\partial f(x_k)}{\partial x^i} \right|^2. \tag{36}$$

We shall show that our hypotheses so far lead to the conclusion that

$$\liminf_{\substack{k \to \infty \\ k \in K}} \alpha_k = 0. \tag{37}$$

Indeed since $\bar{x}$ is not a critical point, there must exist an index i such that either

$$\bar{x}^i > 0 \quad \text{and} \quad \partial f(\bar{x})/\partial x^i \neq 0 \tag{38}$$

or

$$\bar{x}^i = 0 \quad \text{and} \quad \partial f(\bar{x})/\partial x^i < 0. \tag{39}$$

If $i \notin I_k^+$ for an infinite number of indices $k \in K$, then (37) follows from (33), (36), (38), and (39). If $i \in I_k^+$ for an infinite number of indices $k \in K$, then for all those indices we must have $\partial f(x_k)/\partial x^i > 0$, so (39) cannot hold. Therefore, from (38),

$$\bar{x}^i > 0 \quad \text{and} \quad \partial f(\bar{x})/\partial x^i > 0. \tag{40}$$

Since, for all $k \in K$ for which $i \in I_k^+$ [compare with (29)], we have

$$\sum_{j \in I_k^+} \frac{\partial f(x_k)}{\partial x^j} [x_k^j - x_k^j(\alpha_k)] \geq \frac{\partial f(x_k)}{\partial x^i} [x_k^i - x_k^i(\alpha_k)] \geq 0,$$

it follows from (34) and (40) that

$$\lim_{\substack{k \to \infty \\ k \in K}} [x_k^i - x_k^i(\alpha_k)] = 0.$$

Using the above relation, (35), and (40), we obtain (37).

We shall complete the proof by showing that $\{\alpha_k\}_K$ is bounded away from zero thereby contradicting (37). Indeed in view of (31), the subsequences $\{x_k\}_K$, $\{p_k\}_K$, and $\{x_k(\alpha)\}_K$, $\alpha \in [0, 1]$, are uniformly bounded, so by Assumption (A) there exists a scalar $L > 0$ such that, for all $t \in [0, 1]$, $\alpha \in [0, 1]$, and $k \in K$, we have

$$|\nabla f(x_k) - \nabla f[x_k - t[x_k - x_k(\alpha)]]| \leq tL|x_k - x_k(\alpha)|. \tag{41}$$

For all $k \in K$ and $\alpha \in [0, 1]$, we have

$$f[x_k(\alpha)] = f(x_k) + \nabla f(x_k)'[x_k(\alpha) - x_k] + \int_0^1 \{\nabla f(x_k) - \nabla f[x_k - t[x_k - x_k(\alpha)]]\}' \, dt[x_k - x_k(\alpha)],$$

so

$$\begin{aligned} f(x_k) - f[x_k(\alpha)] &= \nabla f(x_k)'[x_k - x_k(\alpha)] \\ &\quad + \int_0^1 \{\nabla f[x_k - t[x_k - x_k(\alpha)]] - \nabla f(x_k)\}' \, dt[x_k - x_k(\alpha)] \\ &\geq \nabla f(x_k)'[x_k - x_k(\alpha)] \\ &\quad - \int_0^1 |\nabla f[x_k - t[x_k - x_k(\alpha)]] - \nabla f(x_k)| \, dt |x_k - x_k(\alpha)|, \end{aligned}$$

and finally, by using (41),

$$f(x_k) - f[x_k(\alpha)] \geq \nabla f(x_k)'[x_k - x_k(\alpha)] - \tfrac{1}{2}L|x_k - x_k(\alpha)|^2. \tag{42}$$

For $i \in I_k^+$, we have $x_k^i(\alpha) = [x_k^i - \alpha p_k^i]^+ \geq x_k^i - \alpha p_k^i$ and $p_k^i > 0$, so $0 \leq x_k^i - x_k^i(\alpha) \leq \alpha p_k^i$. It follows, using (35), that

$$\sum_{i \in I_k^+} |x_k^i - x_k^i(\alpha)|^2 \leq \alpha \sum_{i \in I_k^+} p_k^i [x_k^i - x_k^i(\alpha)] \leq \bar{\lambda}_2 \alpha \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} [x_k^i - x_k^i(\alpha)]. \tag{43}$$

Consider the sets

$$I_{1,k} = \{i \mid \partial f(x_k)/\partial x^i > 0, i \notin I_k^+\}, \qquad I_{2,k} = \{i \mid \partial f(x_k)/\partial x^i \leq 0, i \notin I_k^+\}.$$

For all $i \in I_{1,k}$ we must have $x_k^i > \varepsilon_k$ for otherwise we would have $i \in I_k^+$. Since $|\bar{x} - [\bar{x} - M\nabla f(\bar{x})]^+| \neq 0$, we must have $\liminf_{k\to\infty, k\in K} \varepsilon_k > 0$ and $\varepsilon_k > 0$ for all k. Let $\bar{\varepsilon} > 0$ be such that $\bar{\varepsilon} \leq \varepsilon_k$ for all $k \in K$, and let B be such that $|p_k^i| \leq B$ for all i and $k \in K$. Then, for all $\alpha \in [0, \bar{\varepsilon}/B]$, we have $x_k^i(\alpha) = x_k^i - \alpha p_k^i$, so it follows that

$$\sum_{i\in I_{1,k}} \frac{\partial f(x_k)}{\partial x^i}[x_k^i - x_k^i(\alpha)] = \alpha \sum_{i\in I_{1,k}} \frac{\partial f(x_k)}{\partial x^i} p_k^i \qquad \forall\, \alpha \in \left[0, \frac{\bar{\varepsilon}}{B}\right]. \tag{44}$$

Also, for all $\alpha \geq 0$, we have $x_k^i - x_k^i(\alpha) \leq \alpha p_k^i$, and since $\partial f(x_k)/\partial x^i \leq 0$, for all $i \in I_{2,k}$, we obtain

$$\sum_{i\in I_{2,k}} \frac{\partial f(x_k)}{\partial x^i}[x_k^i - x_k^i(\alpha)] \geq \alpha \sum_{i\in I_{2,k}} \frac{\partial f(x_k)}{\partial x^i} p_k^i. \tag{45}$$

Combining (44) and (45), we obtain

$$\sum_{i\notin I_k^+} \frac{\partial f(x_k)}{\partial x^i}[x_k^i - x_k^i(\alpha)] \geq \alpha \sum_{i\notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i \qquad \forall\, \alpha \in \left[0, \frac{\bar{\varepsilon}}{B}\right]. \tag{46}$$

For all $\alpha \geq 0$, we also have

$$|x_k^i - x_k^i(\alpha)| \leq \alpha |p_k^i| \qquad \forall\, i = 1, \ldots, n.$$

Furthermore, it is easily seen, using Assumption (B), that there exists $\lambda > 0$ such that

$$\sum_{i\notin I_k^+} (p_k^i)^2 \leq \lambda \sum_{i\notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i \qquad \forall\, k \in K.$$

Using the last two relations, we obtain, for all $\alpha \geq 0$,

$$\sum_{i\notin I_k^+} |x_k^i - x_k^i(\alpha)|^2 \leq \alpha^2 \lambda \sum_{i\notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i \qquad \forall\, k \in K. \tag{47}$$

We now combine (42), (43), (46), and (47) to obtain, for all $\alpha \in [0, \bar{\varepsilon}/B]$ with $\alpha \leq 1$ and $k \in K$,

$$\begin{aligned} f(x_k) - f[x_k(\alpha)] \geq{}& \left(\alpha - \frac{\alpha^2 \lambda L}{2}\right) \sum_{i\notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i \\ &+ (1 - \tfrac{1}{2}\alpha \bar{\lambda}_2 L) \sum_{i\in I_k^+} \frac{\partial f(x_k)}{\partial x^i}[x_k^i - x_k^i(\alpha)]. \end{aligned} \tag{48}$$

Suppose α is chosen so that

$$0 \leq \alpha \leq \bar{\varepsilon}/B, \qquad 1 - \tfrac{1}{2}\alpha\lambda L \geq \sigma, \qquad 1 - \tfrac{1}{2}\alpha\bar{\lambda}_2 L \geq \sigma, \qquad \alpha \leq 1, \tag{49}$$

or equivalently

$$0 \leq \alpha \leq \min\left\{\frac{\bar{\varepsilon}}{B}, \frac{2(1-\sigma)}{\lambda L}, \frac{2(1-\sigma)}{\bar{\lambda}_2 L}, 1\right\}. \tag{50}$$

Then we have from (48) and (49), for all $k \in K$,

$$f(x_k) - f[x_k(\alpha)] \geq \sigma\left\{\alpha \sum_{i \notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i + \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i}[x_k^i - x_k^i(\alpha)]\right\}.$$

This means that if (50) is satisfied with $\beta^m = \alpha$, then the inequality (27) of the Armijo-like rule will be satisfied. It follows from the way the stepsize is reduced that α_k satisfies

$$\alpha_k \geq \beta \min\left\{\frac{\bar{\varepsilon}}{B}, \frac{2(1-\sigma)}{\lambda L}, \frac{2(1-\sigma)}{\bar{\lambda}_2 L}, 1\right\} \qquad \forall\, k \in K. \tag{51}$$

This contradicts (37) and proves the proposition. Q.E.D.

We now focus attention at a local minimum x^* satisfying the following second-order sufficiency conditions which are in fact the ones of Proposition 1.31 applied to (SCP), as the reader can easily verify. For all $x \geq 0$, we denote by $A(x)$ the set of indices of active constraints at x; i.e.,

$$A(x) = \{i \mid x^i = 0\} \qquad \forall\, x \geq 0. \tag{52}$$

Assumption (C): *The local minimum x^* of* (SCP) *is such that, for some $\delta > 0$, f is twice continuously differentiable in the open sphere $S(x^*; \delta)$ and there exist positive scalars m_1 and m_2 such that*

$$m_1|z|^2 \leq z'\nabla^2 f(x) z \leq m_2|z|^2 \qquad \forall\, x \in S(x^*; \delta) \quad \text{and} \quad z \neq 0, \quad \text{such that} \quad z^i = 0 \quad \forall\, i \in A(x^*). \tag{53}$$

Furthermore,

$$\partial f(x^*)/\partial x^i > 0 \qquad \forall\, i \in A(x^*). \tag{54}$$

The following proposition demonstrates an important property of the algorithm, namely, that under mild conditions *it is attracted by a local minimum x^* satisfying Assumption* (C) *and identifies the set of active constraints at x^* in a finite number of iterations. Thus, if the algorithm converges to x^*, then after a finite number of iterations it is equivalent to an unconstrained optimization method restricted on the subspace of active constraints at x^*.* This property is instrumental in proving superlinear convergence of the algorithm when the portion of D_k, corresponding to the indices $i \notin I_k^+$, is chosen in a way that approximates the inverse of the portion of the Hessian of f corresponding to these same indices.

Proposition 1.37: Let x^* be a local minimum of (SCP) satisfying Assumption (C), and let Assumption (B) hold in the stronger form whereby, in addition to (31), it is assumed that there exists a scalar $\bar{\lambda}_1 > 0$ such that the diagonal elements d_k^{ii} of the matrices D_k satisfy

$$\bar{\lambda}_1 \leq d_k^{ii} \qquad \forall\, k = 0, 1, \ldots, i \in I_k^+. \tag{55}$$

There exists a scalar $\bar{\delta} > 0$ such that if $\{x_k\}$ is a sequence generated by iteration (25) and for some index $\bar{k}$, we have

$$|x_{\bar{k}} - x^*| \leq \bar{\delta},$$

then $\{x_k\}$ converges to x^*, and we have

$$I_k^+ = A(x_k) = A(x^*) \qquad \forall\, k \geq \bar{k} + 1.$$

Proof: Since f is twice differentiable on $S(x^*; \delta)$, it follows that there exist scalars $L > 0$ and $\delta_1 \in (0, \delta]$ such that for all x and $\bar{x}$ with $|x - x^*| \leq \delta_1$ and $|\bar{x} - x^*| \leq \delta_1$, we have

$$|\nabla f(x) - \nabla f(\bar{x})| \leq L|x - \bar{x}|.$$

Also for x_k sufficiently close to x^*, the scalar

$$w_k = |x_k - [x_k - Mf(x_k)]^+|$$

is arbitrarily close to zero while, in view of (54), we have

$$\left[x_k^i - m^i \frac{\partial f(x_k)}{\partial x^i}\right]^+ = 0 \qquad \forall\, i \in A(x^*),$$

where m^i is the ith diagonal element of M. It follows that, for x_k sufficiently close to x^*, we have

$$x_k^i \leq w_k = \varepsilon_k < \varepsilon \qquad \forall\, i \in A(x^*), \tag{56}$$

while

$$x_k^i > \varepsilon_k \qquad \forall\, i \notin A(x^*). \tag{57}$$

Since, by Assumption (C), $\partial f(x_k)/\partial x^i > 0$ for all $i \in A(x^*)$ and x_k sufficiently close to x^*, (56) and (57) imply that there exists $\delta_2 \in (0, \delta_1]$ such that

$$A(x^*) = I_k^+ \qquad \forall\, k \quad \text{such that} \quad |x_k - x^*| \leq \delta_2. \tag{58}$$

Also there exist scalars $\bar{\varepsilon} > 0$ and $\delta_3 \in (0, \delta_2]$ such that

$$x_k^i > \bar{\varepsilon} \qquad \forall\, i \notin A(x^*) \quad \text{and} \quad k \text{ such that } |x_k - x^*| \leq \delta_3. \tag{59}$$

By repeating the argument in the proof of Proposition 1.36 that led to (51), we find that there exists a scalar $\bar{\alpha} > 0$ such that

$$\alpha_k \geq \bar{\alpha} \qquad \forall\, k \quad \text{such that} \quad |x_k - x^*| \leq \delta_3. \tag{60}$$

By using (55) and (58), it follows that

$$0 < \bar{\lambda}_1 \, \partial f(x_k)/\partial x^i \leq p_k^i \qquad \forall\, i \in A(x^*) \quad \text{and} \quad k \text{ such that } |x_k - x^*| \leq \delta_3, \tag{61}$$

while, by Assumption (B), there exists a scalar $\lambda > 0$ such that

$$\sum_{i \notin A(x^*)} |p_k^i|^2 \leq \lambda \sum_{i \notin A(x^*)} \left| \frac{\partial f(x_k)}{\partial x^i} \right|^2 \qquad \forall\, k \quad \text{such that} \quad |x_k - x^*| \leq \delta_3. \tag{62}$$

Since $\partial f(x^*)/\partial x^i > 0$ for all $i \in A(x^*)$ and $\partial f(x^*)/\partial x^i = 0$ for all $i \notin A(x^*)$, it follows from (58)–(62) that there exists a scalar $\delta_4 \in (0, \delta_3]$ such that

$$A(x^*) = A(x_{k+1}) \qquad \forall\, k \quad \text{such that} \quad |x_k - x^*| \leq \delta_4 \tag{63}$$

and

$$|x_{k+1} - x^*| \leq \delta_3 \qquad \forall\, k \quad \text{such that} \quad |x_k - x^*| \leq \delta_4. \tag{64}$$

In view of (58), we obtain, from (63) and (64),

$$A(x^*) = A(x_{k+1}) = I_{k+1}^+ \qquad \forall\, k \quad \text{such that} \quad |x_k - x^*| \leq \delta_4. \tag{65}$$

Thus when $|x_k - x^*| \leq \delta_4$, we have $|x_{k+1} - x^*| \leq \delta_3$, $A(x^*) = A(x_{k+1})$, and the $(k+1)$th iteration of the algorithm reduces to an iteration of an unconstrained minimization algorithm on the subspace of active constraints at x^* to which Proposition 1.12 applies. From this proposition, it follows that there exists an open set $N(x^*)$ containing x^* such that $N(x^*) \subset S(x^*; \delta_4)$ and with the property that if $x_{k+1} \in N(x^*)$ and $A(x_{k+1}) = A(x^*)$, then $x_{k+2} \in N(x^*)$ and, by (63), $A(x_{k+2}) = A(x^*)$. This argument can be repeated and shows that if for some $\bar{k} \geq 0$ we have

$$x_{\bar{k}} \in N(x^*), \qquad A(x_{\bar{k}}) = A(x^*),$$

then $\{x_k\} \to x^*$ and

$$x_k \in N(x^*), \qquad A(x_k) = A(x^*) \qquad \forall\, k \geq \bar{k}.$$

To complete the proof, it is sufficient to show that there exists $\bar{\delta} > 0$ such that if $|x_k - x^*| \leq \bar{\delta}$ then $x_{k+1} \in N(x^*)$ and $A(x_{k+1}) = A(x^*)$. Indeed by repeating the argument that led to (63) and (64), we find that given any $\tilde{\delta} > 0$ there exists a $\bar{\delta} > 0$ such that if $|x_k - x^*| \leq \bar{\delta}$, then

$$|x_{k+1} - x^*| \leq \tilde{\delta}, \qquad A(x_{k+1}) = A(x^*).$$

By taking $\tilde{\delta}$ sufficiently small so that $S(x^*; \tilde{\delta}) \subset N(x^*)$ the proof is completed. Q.E.D.

Under the assumptions of Proposition 1.37, we see that if the algorithm converges to a local minimum x^* satisfying Assumption (C) then it reduces eventually to an unconstrained minimization method restricted to the subspace

$$S^* = \{x \mid x^i = 0, \forall\, i \in A(x^*)\}.$$

Furthermore, as shown in the proof of Proposition 1.37 [compare with (58)], for some index $\bar{k}$, we shall have

$$I_k^+ = A(x^*) \qquad \forall\, k \geq \bar{k}. \tag{66}$$

This shows that if the portion of the matrix D_k corresponding to the indices $i \notin I_k^+$ is chosen to be the inverse of the Hessian of f with respect to the indices $i \notin I_k^+$, then the algorithm eventually reduces to Newton's method restricted to the subspace S^*.

More specifically, by rearranging indices if necessary, assume without loss of generality that

$$I_k^+ = \{r_k + 1, \ldots, n\}, \tag{67}$$

where r_k is some integer. Then D_k has the form

$$D_k = \begin{bmatrix} \bar{D}_k & \vdots & 0 \\ \cdots & \vdots & \cdots \\ & \vdots & d^{r_k+1} \quad\quad 0 \\ & \vdots & \quad\quad\ddots \\ & \vdots & 0 \quad\quad d^n \end{bmatrix}, \tag{68}$$

where $d_k^i > 0$, $i = r_k + 1, \ldots, n$, and $\bar{D}_k$ can be an *arbitrary* positive definite matrix. Suppose we choose $\bar{D}_k$ to be the inverse of the Hessian of f with respect to the indices $i = 1, \ldots, r_k$; i.e., the elements $[\bar{D}_k^{-1}]_{ij}$ of $\bar{D}_k^{-1}$ are

$$[\bar{D}_k^{-1}]_{ij} = \partial^2 f(x_k)/\partial x^i\, \partial x^j \qquad \forall\, i, j \notin I_k^+. \tag{69}$$

By Assumption (C), $\nabla^2 f(x^*)$ is positive definite on S^*, so it follows from (66) that this choice is well defined and satisfies the assumption of Proposition 1.37 for k sufficiently large. Since the conclusion of this proposition asserts that the method eventually reduces to Newton's method restricted to the subspace S^*, a superlinear convergence rate result follows. This type of argument can be used to construct a number of Newton-like and quasi-Newton methods and prove corresponding convergence and rate of convergence results. We state one of the simplest such results regarding a Newton-like algorithm which is well suited for problems where f is strictly convex and twice differentiable. Its proof follows simply from the preceding discussion and Propositions 1.15 and 1.17 and is left to the reader.

Proposition 1.38: Let f be convex and twice continuously differentiable. Assume that (SCP) has a unique optimal solution x^* satisfying Assumption (C), and there exist positive scalars m_1 and m_2, such that

$$m_1|z|^2 \leq z'\nabla^2 f(x)z \leq m_2|z|^2 \qquad \forall\, z \in \{x \mid f(x) \leq f(x_0)\}.$$

Assume also that in the algorithm (22)–(27), the matrix D_k is given by $D_k = H_k^{-1}$, where H_k is the matrix with elements H_k^{ij} given by

$$H_k^{ij} = \begin{cases} 0 & \text{if } i \neq j \text{ and either } i \in I_k^+ \text{ or } j \in I_k^+, \\ \partial^2 f(x_k)/\partial x^i\, \partial x^j & \text{otherwise.} \end{cases}$$

Then the sequence $\{x_k\}$ generated by iteration (25) converges to x^*, and the rate of convergence of $\{|x_k - x^*|\}$ is superlinear (of order at least two if $\nabla^2 f$ is Lipschitz continuous in a neighborhood of x^*).

It is worth noting that when $f(x)$ is a positive definite quadratic function, the algorithm of Proposition 1.38 finds the unique solution x^* in a finite number of iterations, assuming x^* satisfies Assumption (C).

An additional property of the algorithm of Proposition 1.38 is that after a finite number of iterations and once the set of binding constraints is identified, the initial unity stepsize is accepted by the Armijo rule. Computational experience with the algorithm suggests that this is also true for most iterations even before the set of binding constraints is identified. In some cases, however, it may be necessary to reduce the initial unity stepsize several times before a sufficient reduction in objective function value is effected. A typical situation where this may occur is when the scalar $\tilde{\gamma}_k$ defined by

$$\tilde{\gamma}_k = \min\{1, \tilde{\alpha}_k\}, \qquad \tilde{\alpha}_k = \sup\{\alpha \mid x_k^i - \alpha p_k^i \geq 0, x_k^i > 0, i \notin I_k^+\}$$

is much smaller than unity. Under these circumstances a nonbinding constraint that was not included in the set I_k^+ becomes binding after a small movement along the arc $\{x_k(\alpha) \mid \alpha \geq 0\}$ and it may happen that the objective function value increases as α becomes larger than $\tilde{\gamma}_k$. To correct such a situation, it may be useful to modify the Armijo rule so that if after a fixed number r of trial stepsizes $1, \beta, \ldots, \beta^{r-1}$ have failed to pass the Armijo rule test, then $\tilde{\gamma}_k$ is computed and, if it is smaller than β^{r-1}, it is used as the next trial stepsize.

Another (infrequent) situation, where the algorithm of Proposition 1.38 can exhibit a large number of stepsize reductions and slow convergence when far from the optimum, arises sometimes if the set of indices

$$\tilde{I}_k^+ = \{i \mid 0 \leq x_k^i \leq \varepsilon_k, p_k^i > 0\}, \tag{70}$$

where $p_k = D_k \nabla f(x_k)$, is strictly larger than the set I_k^+ of (22). (Note that, under the assumptions of Proposition 1.38, we always have $I_k^+ \subset \tilde{I}_k^+$ with equality holding in a neighborhood of the optimal solution x^*.) Under these circumstances, the initial motion along the arc $\{x(\alpha) | \alpha \geq 0\}$ may be along a search direction that is not a Newton direction on any subspace. A possible remedy for this difficulty is to combine the Armijo rule with some form of line minimization rule.

Extension to Upper and Lower Bounds

The algorithm (22)–(27) described so far in this section can be easily extended to handle problems of the form

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad b_1 \leq x \leq b_2, \end{aligned}$$

where b_1 and b_2 are given vectors of lower and upper bounds. The set I_k^+ is replaced by

$$\begin{aligned} I_k^{\#} = \{ i | & b_1^i \leq x_k^i \leq b_1^i + \varepsilon_k \text{ and } \partial f(x_k)/\partial x^i > 0 \\ & \text{or } b_2^i - \varepsilon_k \leq x_k^i \leq b_2^i \text{ and } \partial f(x_k)/\partial x^i < 0 \}, \end{aligned}$$

and the definition of $x_k(\alpha)$ is changed to

$$x_k(\alpha) = [x_k - \alpha D_k \nabla f(x_k)]^{\#},$$

where for all $z \in R^n$ we denote by $[z]^{\#}$ the vector with coordinates

$$[z]^{\#i} = \begin{cases} b_2^i & \text{if} \quad b_2^i \leq z^i, \\ z^i & \text{if} \quad b_1^i < z^i < b_2^i, \\ b_1^i & \text{if} \quad z^i \leq b_1^i. \end{cases}$$

The scalar ε_k is given by

$$\varepsilon_k = \min\{\varepsilon, | x_k - [x_k - M \nabla f(x_k)]^{\#} | \}.$$

The matrix D_k is positive definite and diagonal with respect to $I_k^{\#}$, and M is a fixed diagonal positive definite matrix. The iteration is given by

$$x_{k+1} = x_k(\alpha_k),$$

where α_k is chosen by the Armijo rule (26), (27) with $[x_k^i - x_k^i(\beta^m)]^+$ replaced by $[x_k^i - x_k^i(\beta^m)]^{\#}$.

Similar extensions of the basic algorithm can be provided for problems where only some of the variables x^i are simply constrained by upper and/or lower bounds.

1.6 Notes and Sources

Notes on Section 1.2: The proof of the second implicit function theorem may be found in Hestenes (1966, p. 23). The theorem itself is apparently due to Bliss (Hestenes, personal communication).

Notes on Section 1.3: The convergence analysis of gradient methods given here stems from the papers of Goldstein (1962, 1966), and bears similarity with the corresponding analysis in Ortega and Rheinboldt (1970). Some other influential works in this area are Armijo (1966), Wolfe (1969), and Daniel (1971). Zangwill (1969) and Polak (1971) have proposed general convergence theories for optimization algorithms. The gradient method with constant stepsize was first analyzed by Poljak (1963). Proposition 1.12 is thought to be new. The linear convergence rate results stem from Kantorovich (1945) and Poljak (1963), while the superlinear rate results stem from Goldstein and Price (1967). For convergence rate analysis of the steepest descent method near local minima with singular Hessian, see Dunn (1981b). The spacer step theorem (Proposition 1.16) is due to Zangwill (1969). For an extensive analysis and references on Newton-like methods, see Ortega and Rheinboldt (1970). The modification scheme for Newton's method based on the Cholesky factorization is related to one due to Murray (1972).

Conjugate direction methods were originally developed in Hestenes and Stiefel (1952). Extensive presentations may be found in Faddeev and Faddeeva (1963), Luenberger (1973), and Hestenes (1980). Scaled $(k+1)$-step conjugate gradient methods for problems with Hessian matrix of the form

$$Q = M + \sum_{i=l}^{k} v_i v_i'$$

were first proposed in Bertsekas (1974a). For further work on this subject, see Oren (1978).

Extensive surveys of quasi-Newton methods can be found in Avriel (1976), Broyden (1972), and Dennis and Moré (1977).

Notes on Section 1.4: Presentations of optimality conditions for constrained optimization can be found in many sources including Fiacco and McCormick (1968), Mangasarian (1969), Cannon *et al.* (1970), Luenberger (1973), and Avriel (1976). For a development of optimality conditions based on the notion of augmentability, which is intimately related to methods of multipliers, see Hestenes (1975).

Notes on Section 1.5: The methods in this section are new and were developed while the monograph was being written. Extensions to general

linear constraints may be found in Bertsekas (1980c). The methods are particularly well suited for large scale problems with many simple constraints. An example is nonlinear multicommodity flow problems arising in communication and transportation networks (see Bertsekas and Gafni, 1981). The constrained version of the Armijo rule (26), (27) is based on a similar rule first proposed in Bertsekas (1974c). The main advantage that the methods of this section offer over methods based on active set strategies [compare with Gill and Murray (1974) and Ritter (1973)] is that there is no limit to the number of constraints that can be added or dropped from the active set in a single iteration, and this is significant for problems of large dimension. At the same time, there is no need to solve a quadratic programming problem at each iteration as in the Newton and quasi-Newton methods of Levitin and Poljak (1965), Garcia-Palomares (1975), and Brayton and Cullum (1979).

References

The following abbreviations have been used in the reference list.

M.P.	Math. Programming	S.J.C.	SIAM J. on Control
M.S.	Management Science	S.J.C.O.	SIAM J. on Control and Optimization
J.O.T.A.	J. Opt. Th. & Appl.	O.R.	Operations Research

Armijo, L. (1966). Minimization of functions having continuous partial derivatives. *Pacific J. Math.* **16**, 1–3.

Arrow, K. J., and Solow, R. M. (1958). Gradient methods for constrained maxima with weakened assumptions. In "Studies in Linear and Nonlinear Programming" (K. J. Arrow, L. Hurwitz, and H. Uzawa, eds.), pp. 166–176. Stanford Univ. Press, Stanford, California.

Arrow, K. J., Hurwicz, L., and Uzawa, H., eds. (1958). "Studies in Linear and Nonlinear Programming." Stanford Univ. Press, Stanford, California.

Arrow, K. J., Gould, F. J., and Howe, S. M. (1973). A general saddle point result for constrained minimization. *M.P.* **5**, 225–234.

Aubin, J. P., and Ekeland, I. (1976). Estimates of the duality gap in nonconvex optimization. *Math. Oper. Res.* **1**, 225–245.

Auslender, A. (1976). "Optimization: Methodes Numeriques." Mason, Paris.

Avriel, M. (1976). "Nonlinear Programming: Analysis and Methods." Prentice-Hall, Englewood Cliffs, New Jersey.

Bazaraa, M. S., and Goode, J. J. (1979). "An Extension of Armijo's Rule to Minimax and Quasi-Newton Methods for Constrained Optimization," Indust. Systems Engr. Rep. Georgia Inst. of Tech., Atlanta, Georgia. *European J. of Operational Research*, to appear.

Beale, E. M. L. (1972). A derivation of conjugate gradients. *In* "Numerical Methods for Nonlinear Optimization" (F. A. Lootsma, ed.), pp. 39–43. Academic Press, New York.

Bertsekas, D. P. (1973). Convergence rate of penalty and multiplier methods. *Proc. 1973 IEEE Confer. Decision Control, San Diego, Calif.*, pp. 260–264.

Bertsekas, D. P. (1974a). Partial conjugate gradient methods for a class of optimal control problems. *IEEE Trans Automat. Control* **19**, 209–217.

Bertsekas, D. P. (1974b). Nondifferentiable optimization via approximation. *Proc. Annual Allerton Confer. Circuit System Theory, 12th, Allerton Park, Ill.* pp. 41–52. Also *in* "Mathematical Programming Study 3" (M. Balinski and P. Wolfe, eds.). pp. 1–25. North-Holland Publ., Amsterdam, 1975.

Bertsekas, D. P. (1974c). On the Goldstein–Levitin–Poljak gradient projection method. *Proc. 1974 IEEE Decision Control Conf., Phoenix, Ariz.*, pp. 47–52. Also *IEEE Trans. Automat. Control* **21**, 174–184 (1976).

Bertsekas, D. P. (1975a). Multiplier methods for convex programming. *IEEE Trans. Automat. Control* **20**, 385–388.

Bertsekas, D. P. (1975b). Necessary and sufficient conditions for a penalty method to be exact. *M.P.* **9**, 87–99.

Bertsekas, D. P. (1975c). Combined primal-dual and penalty methods for constrained minimization. *S.J.C.* **13**, 521–544.

Bertsekas, D. P. (1976a). On penalty and multiplier methods for constrained optimization. *S.J.C.O.* **14**, 216–235.

Bertsekas, D. P. (1976b). Multiplier methods: A survey. *Automatica—J. IFAC* **12**, 133–145.

Bertsekas, D. P. (1976c). Newton's method for linear optimal control problems. *Proc. IFAC Symp. Large-Scale Systems, Udine, Italy* pp. 353–359.

Bertsekas, D. P. (1976d). Minimax methods based on approximation. *Proc. 1976 Johns Hopkins Confer. Inform. Sci. Systems, Baltimore, Md.* pp. 363–365.

Bertsekas, D. P. (1976e). A new algorithm for solution of resistive networks involving diodes. *IEEE Trans. Circuits and Systems* **23**, 599–608.

Bertsekas, D. P. (1977). Approximation procedures based on the method of multipliers. *J.O.T.A.* **23**, 487–510.

Bertsekas, D. P. (1978). On the convergence properties of second order methods of multipliers. *J.O.T.A* **25**, 443–449.

Bertsekas, D. P. (1979a). A convergence analysis of the method of multipliers for nonconvex constrained optimization. Presented at *Proc. Workshop Augmented Lagrangians, IIASA, Vienna.*

Bertsekas, D. P. (1979b) Convexification procedures and decomposition algorithms for large-scale nonconvex optimization problems. *J.O.T.A.* **29**, 169–197.

Bertsekas, D. P. (1980a). "Enlarging the region of convergence of Newton's method for constrained optimization," LIDS Rep. R-985. MIT, Cambridge, Massachusetts. Also *J.O.T.A.* **36**, (1982) 221–252.

Bertsekas, D. P. (1980b). Variable metric methods for constrained optimization based on differentiable exact penalty functions. *Proc. Allerton Confer. Comm. Control Comput., Allerton Park, Ill.* pp. 584–593.

Bertsekas, D. P. (1980c). "Projected Newton methods for optimization problems with simple constraints," LIDS Rep. R-1025. MIT, Cambridge, Massachusetts. Also *S.J.C.O.* **20**, (1982), pp. 221–246.

Bertsekas, D. P., and Gafni, E. (1983). "Projected Newton Methods and Optimization of Multicommodity Flows," *IEEE Trans. Automat. Control*, AC-28, pp. 1090–1096.

Bertsekas, D. P., and Mitter, S. K. (1971). Steepest descent for optimization problems with nondifferentiable cost functionals. *Proc. Annual Princeton Confer. Inform. Sci. Systems, 5th, Princeton, N.J.* pp. 347–351.

Bertsekas, D. P., and Mitter, S. K. (1973). A descent numerical method for optimization problems with nondifferentiable cost functionals. *S.J.C.* **11**, 637–652.

Bertsekas, D. P., Lauer, G. S., Sandell, N. R., Jr., and Posbergh, T. A. (1981). Optimal short term scheduling of large-scale power systems. *Proc. 1981 IEEE Confer. Decision Control, San Diego, Calif.* pp. 432-443, *IEEE Trans. Automat. Control*, Vol. AC-28, 1983, pp. 1–11.

Biggs, M. C. (1972). Constrained minimization using recursive equality quadratic programming. *In* "Numerical Methods for Nonlinear Optimization" (F. A. Lootsma, ed.), pp. 411–428. Academic Press, New York.

Biggs, M. C. (1978). On the convergence of some constrained minimization algorithms based on recursive quadratic programming. *J. Inst. Math. Appl.* **21**, 67–81.

Boggs, P. T., and Tolle, J. W. (1980). Augmented Lagrangians which are quadratic in the multiplier. *J.O.T.A.* **31**, 17–26.

Boggs, P. T., Tolle, J. W., and Wang, P. (1982). On the local convergence of quasi-Newton methods for constrained optimization. *S.J.C.O.* **20**, 161–171.

Brayton, R. K., and Cullum, J. (1979). An algorithm for minimizing a differentiable function subject to box constraints and errors. *J.O.T.A.* **29**, 521–558.

Brent, R. P. (1972). "Algorithms for Minimization without Derivatives." Prentice-Hall, Engelwood Cliffs, New Jersey.

Broyden, C. G. (1970). The convergence of a class of double rank minimization algorithms. *J. Inst. Math. Appl.* **6**, 76–90.

Broyden, C. G. (1972). Quasi-Newton methods. *In* "Numerical Methods for Unconstrained Optimization" (W. Murray, ed.), pp. 87–106. Academic Press, New York.

Broyden, C. G., Dennis, J., and Moré, J. J. (1973). On the local and superlinear convergence of quasi-Newton methods. *J. Inst. Math. Appl.* **12**, 223–245.

Brusch, R. B. (1973). A rapidly convergent method for equality constrained function minimization. *Proc. 1973 IEEE Confer. Decision Control, San Diego, Calif.* pp. 80–81.

Buys, J. D. (1972). Dual algorithms for constrained optimization. Ph.D. Thesis, Rijksuniversiteit de Leiden.

Campos-Filho, A. S. (1971). Numerical computation of optimal control sequences. *IEEE Trans. Automat. Control* **16**, 47–49.

Cannon, M. D., Cullum, C. D., and Polak, E. (1970). "Theory of Optimal Control and Mathematical Programming," McGraw-Hill, New York.

Chamberlain, R. M. (1979). Some examples of cycling in variable metric methods for constrained minimization. *M.P.* **16**, 378–383.

Chamberlain, R. M. (1980). The theory and application of variable metric methods to constrained optimization problems. Ph.D. Thesis, Univ. of Cambridge, Cambridge, England.

Chamberlain, R. M., Lemarechal, C., Pedersen, H. C., and Powell, M. J. D. (1979). The watchdog technique for forcing convergence in algorithms for constrained optimization. *Internat. Symp. Math. Programming, 10th, Montreal, Math. Programming Stud.* **16**, (1982), to appear.

Coleman, T. F., and Conn, A. R. (1980a). "Nonlinear Programming via an Exact Penalty Function: Asymptotic Analysis," Comput. Sci. Dept., Rep. CS-80-30. Univ. of Waterloo, *M.P.*, to appear.

Coleman, T. F., and Conn, A. R. (1980b). "Nonlinear Programming via an Exact Penalty Function: Global Analysis," Comput. Sci. Dept., Rep. CS-80-31. Univ. of Waterloo, *M.P.*, to appear.

Daniel, J. W. (1971). "The Approximate Minimization of Functionals." Prentice-Hall, Englewood Cliffs, New Jersey.

Davidon, W. C. (1959). "Variable Metric Method for Minimization," R and D Rep. ANL-599 (Ref.). U.S. At. Energy Commission, Argonne Nat. Lab., Argonne, Illinois.

Dembo, R. S., Eisenstadt, S. C., and Steihaug, T. (1980). "Inexact Newton Methods," Working Paper, Ser. B, No. 47. Yale Sch. Organ. Management, New Haven, Connecticut.

Dennis, J. E., and Moré, J. J. (1974). A characterization of superlinear convergence and its application to quasi-Newton methods. *MC* **28**, 549–560.

Dennis, J. E., and Moré, J. J. (1977). Quasi-Newton methods: motivation and theory. *SIAM Rev.* **19**, 46–89.

DiPillo, G., and Grippo, L. (1979a). A new class of augmented Lagrangians in nonlinear programming. *S.J.C.O.* **17**, 618–628.

DiPillo, G., and Grippo, L. (1979b). "An Augmented Lagrangian for Inequality Constraints in Nonlinear Programming Problems," Rep. 79–22. Ist. Automat., Univ. di Roma.

DiPillo, G., Grippo, L., and Lampariello, F. (1979). A method for solving equality constrained optimization problems by unconstrained minimization. *Proc. IFIP Confer. Optim. Tech., 9th, Warsaw* pp. 96–105.

Dixon, L. C. W. (1972a). Quasi-Newton algorithms generate identical points. *M.P.* **2**, 383–397.

Dixon, L. C. W. (1972b). Quasi-Newton algorithms generate identical points. *M.P.* **3**, 346–358.

Dixon, L. C. W. (1980). "On the Convergence Properties of Variable Metric Recursive Quadratic Programming Methods," Numer. Optim. Centre, Rep. No. 110. Hatfield Polytechnic, Hatfield, England.

Dolecki, S., and Rolewicz, S. (1979). Exact penalties for local minima. *S.J.C.O.* **17**, 596–606.

Dunn, J. C. (1980). Newton's method and the Goldstein Step Length Rule for constrained minimization problems. *S.J.C.O.* **18**, 659–674.

Dunn, J. C. (1981). Global and asymptotic convergence rate estimates for a class of projected gradient processes. *S.J.C.O.* **19**, 368–400.

Ekeland, I., and Teman, R. (1976). "Convex Analysis and Variational Problems." North-Holland Publ., Amsterdam.

Ermoliev, Y. M., and Shor, N. Z. (1967). On the minimization of nondifferentiable functions." *Kibernetika (Kiev)* **3**, 101–102.

Evans, J. P., Gould, F. J., and Tolle, J. W. (1973). Exact penalty functions in nonlinear programming. *Math. Programming* **4**, 72–97.

Everett, H. (1963). Generalized Lagrange multiplier method for solving problems of optimal allocation of resources. *O.R.* **11**, 399–417.

Fadeev, D. K., and Fadeeva, V. N. (1963). "Computational Methods of Linear Algebra." Freeman, San Francisco, California.

Fiacco, A. V., and McCormick, G. P. (1968). "Nonlinear Programming: Sequential Unconstrained Minimization Techniques." Wiley, New York.

Fletcher, R. (1970). A class of methods for nonlinear programming with termination and convergence properties. *In* "Integer and Nonlinear Programming" (J. Abadie, ed.), pp. 157–173. North-Holland Publ., Amsterdam.

Fletcher, R. (1973). A class of methods for nonlinear programming: III. Rates of convergence. *In* "Numerical Methods for Nonlinear Optimization" (F. A. Lootsma, ed.), pp. 371–381. Academic Press, New York.

Fletcher, R. (1975). An ideal penalty function for constrained optimization. *In* "Nonlinear Programming 2" (O. Mangasarian, R. Meyer, and S. Robinson, eds.) pp. 121–163. Academic Press, New York.

Fletcher, R., and Freeman, T. L. (1977). A modified Newton method for minimization, *J.O.T.A.* **23**, 357–372.

Fletcher, R., and Lill, S. (1971). A class of methods for nonlinear programming: II. computational experience. In "Nonlinear Programming" (J. B. Rosen, O. L. Mangasarian, and K. Ritter, eds.), pp. 67–92. Academic Press, New York.

Fletcher, R., and Powell, M. J. D. (1963). A rapidly convergent descent algorithm for minimization. *Comput. J.* **6**, 163–168.

Fletcher, R., and Reeves, C. M. (1964). Function minimization by conjugate gradients. *Comput. J.* **7**, 149–154.

Gabay, D. (1979). Reduced quasi-Newton methods with feasibility improvement for nonlinearly constrained optimization. *Math. Programming Stud.* **16**, (1982), to appear.

Garcia-Palomares, U. M. (1975). Superlinearly convergent algorithms for linearly constrained optimization. *In* "Nonlinear Programming 2" (O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds.), pp. 101–119. Academic Press, New York.

Garcia-Palomares, U. M., and Mangasarian, O. L. (1976). Superlinearly convergent quasi-Newton algorithms for nonlinearly constrained optimization problems. *M.P.* **11**, 1–13.

Geoffrion, A. (1974). Lagrangian relaxation for integer programming. *Math. Programming Stud.* **2**, 82–114.

Gill, P. E., and Murray, W. (1972). Quasi-Newton methods for unconstrained optimization. *J. Inst. Math. Appl.* **9**, 91–108.

Gill, P. E., and Murray, W. (1974). "Numerical Methods for Constrained Optimization." Academic Press, New York.

Gill, P. E., Murray, W., and Wright, M. H. (1981). "Practical Optimization." Academic Press, New York.

Glad, T. (1979). Properties of updating methods for the multipliers in augmented Lagrangians. *J.O.T.A.* **28**, 135–156.

Glad, T., and Polak, E. (1979). A multiplier method with automatic limitation of penalty growth. *M.P.* **17**, 140–155.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Math. Comp.* **24**, 23–26.

Goldfarb, D. (1980). Curvilinear path steplength algorithms for minimization which use directions of negative curvature. *M.P.* **18**, 31–40.

Goldstein, A. A. (1962). Cauchy's method of minimization. *Numer. Math.* **4**, 146–150.

Goldstein, A. A. (1964). Convex programming in Hibert Space. *Bull. Amer. Math. Soc.* **70**, 709–710.

Goldstein, A. A. (1966). Minimizing functionals on normed linear spaces. *S.J.C.* **4**, 81–89.

Goldstein, A. A. (1974). On gradient projection. *Proc. Annual Allerton Confer., 12th, Allerton Park, Ill.* pp. 38–40.

Goldstein, A. A. (1977). Optimization of Lipschitz continuous functions. *M.P.* **13**, 14–22.

Goldstein, A. A., and Price, J. B. (1967). An effective algorithm for minimization. *Numer. Math.* **10**, 184–189.

Golshtein, E. G. (1972). A generalized gradient method for finding saddle points. *Matecon* **8**, 36–52.

Greenstadt, J. (1970). Variations on variable metric methods. *Math. Comput.* **24**, 1–18.

Haarhoff, P. C., and Buys, J. D. (1970). A new method for the optimization of a nonlinear function subject to nonlinear constraints. *Comput. J.* **13**, 178–184.

Han, S. P. (1977a). Dual variable metric algorithms for constrained optimization. *S.J.C.O.* **15**, 135–194.

Han, S. P. (1977b). A globally convergent method for nonlinear programming. *J.O.T.A.* **22**, 297–309.

Han S. P., and Mangasarian, O. L. (1979). Exact penalty functions in nonlinear programming. *M.P.* **17**, 251–269.

Han, S. P., and Mangasarian, O. L. (1981). A dual differentiable exact penalty function. Computer Sciences Tech. Rep. #434, Univ. of Wisconsin, Madison.

Held, M., and Karp, R. M. (1970). The Traveling Salesman Problem and minimum spanning trees. *O.R.* **18**, 1138–1162.

Held, M., Wolfe, P., and Crowder, H. (1974). Validation of subgradient optimization. *M.P.* **6**, 62–88.

Hestenes, M. R. (1966). "Calculus of Variations and Optimal Control Theory." Wiley, New York.

Hestenes, M. R. (1969). Multiplier and gradient methods. *J.O.T.A.* **4**, 303–320.

Hestenes, M. R. (1975). "Optimization Theory: The Finite Dimensional Case." Wiley, New York.

Hestenes, M. R. (1980). "Conjugate Direction Methods in Optimization." Springer-Verlag, Berlin and New York.

Hestenes, M. R., and Stiefel, E. L. (1952). Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Standards Sect. B* **49**, 409–436.

Howe, S. (1973). New conditions for exactness of a simple penalty function. *S.J.C.* **11**, 378–381.

Jijtontrum, K. (1980). Accelerated convergence for the Powell/Hestenes multiplier methods. *M.P.* **18**, 197–214.

Kantorovich, L. V. (1945). On an effective method of solution of extremal problems for a quadratic functional. *Dokl. Akad. Nauk SSSR* **48**, 483–487.

Klessig, R., and Polak, E. (1972). Efficient implementation of the Polak–Ribiere conjugate gradient algorithms. *S.J.C.* **10**, 524–549.

Korpelevich, G. M. (1976). The extragradient method for finding saddle points and other problems. *Matecon* **12**, 35–49.

Kort, B. W. (1975a). Combined primal-dual and penalty function algorithms for nonlinear programming. Ph.D. Thesis, Stanford Univ., Stanford. California.

Kort, B. W. (1975b). Rate of convergence of the method of multipliers with inexact minimization. *In* "Nonlinear Programming 2" (O. Mangasarian, R. Meyer, and S. Robinson, eds.), pp. 193–214. Academic Press, New York.

Kort, B. W., and Bertsekas, D. P. (1972). A new penalty function method for constrained minimization. *Proc 1972 IEEE Confer. Decision Control, New Orleans. La.* pp. 162–166.

Kort, B. W., and Bertsekas, D. P. (1973). Multiplier methods for convex programming. *Proc. 1073 IEEE Conf. Decision Control, San Diego, Calif.* pp. 428–432.

Kort, B. W., and Bertsekas, D. P. (1976). Combined primal-dual and penalty methods for convex programming. *S.J.C.O.* **14**, 268–294.

Kwakernaak, H., and Strijbos, R. C. W. (1972). Extremization of functions with equality constraints. *M.P.* **2**, 279–295.

Lasdon, L. S. (1970). "Optimization Theory for Large Systems." Macmillan, New York.

Lauer, G., Bertsekas, D. P., Sandell, N., and Posbergh, T. (1981). Optimal solution of large-scale unit commitment problems. *IEEE Trans. Power Systems Apparatus.* **101**, 79–86.

Lemarechal, C. (1974). An algorithm for minimizing convex functions." *In* "Information Processing '74" (J. L. Rosenfeld, ed.), pp. 552–556. North-Holland Publ., Amsterdam.

Lemarechal, C. (1975). An extension of Davidon methods to nondifferentiable problems. *Math. Programming Stud.* **3**, 95–109.

Lenard, M. L. (1973). Practical convergence conditions for unrestrained optimization. *M.P.* **4**, 309–325.

Lenard, M. L. (1976). Convergence conditions for restarted conjugate gradient methods with inaccurate line searches. *M.P.* **10**, 32–51.

Lenard, M. L. (1979). A computational study of active set strategies in nonlinear programming with linear constraints. *M.P.* **16**, 81–97.

Levitin, E. S., and Poljak, B. T. (1965). Constrained minimization methods. *Z. Vyčisl. Mat. i Mat. Fiz.* **6**. 787–823. Engl. transl. in *C.M.M.P.* **6**, 1–50 (1966).

Luenberger, D. G. (1979). "Optimization by Vector Space Methods." Wiley, New York.

Luenberger, D. G. (1970). Control problems with kinks. *IEEE Trans. Automat. Control* **15**, 570–575.

Luenberger, D. G. (1973). "Introduction to Linear and Nonlinear Programming." Addison-Wesley, Reading, Massachusetts.

Luque, J. R. (1981). "Asymptotic convergence analysis of the proximal point algorithm." LIDS Rep. P-1142, M.I.T., Cambridge, Mass. To appear in *S.J.C.O.*

McCormick, G. P. (1969). Anti-zig-zagging by bending. *M.S.* **15**, 315–319.

McCormick, G. P. (1978). An idealized exact penalty function. *In* "Nonlinear Programming 3" (O. Mangasarian, R. Meyer, and S. Robinson, eds.) pp. 165–195. Academic Press, New York.

McCormick, G. P., and Ritter, K. (1972). Methods of conjugate directions versus quasi-Newton methods. *M.P.* **3**, 101–116.

Magnanti, T. L. Shapiro, J. F., and Wagner, M. H. (1976). Generalized linear programming solves the dual. *M.S.* **22**, 1195–1203.

Maistrovskii, G. D. (1976). Gradient methods for finding saddle points. *Matecon* **12**. 3–22.

Mangasarian, O. L. (1969). "Nonlinear Programming." Prentice-Hall, Englewood Cliffs. New Jersey.

Mangasarian, O. L. (1974). Unconstrained methods in optimization. *Proc. Allerton Conf. Circuit System Theory, 12th, Univ. Ill., Urbana* pp. 153–160.

Mangasarian, O. L. (1975). Unconstrained Lagrangians in nonlinear programming. *S.J.C.* **13**, 772–791.

Maratos, N. (1978). Exact penalty function algorithms for finite dimensional and control optimization problems. Ph.D. Thesis, Imperial College Sci. Tech., Univ. of London.

Mayne, D. Q., and Maratos, N. (1979). A first order exact penalty function algorithm for equality constrained optimization problems. *M.P.* **16**, 303–324.

Mayne, D. Q., and Polak, E. (1978). "A Superlinearly Convergent Algorithm for Constrained Optimization Problems," Res. Rep. 78–52. Dept. Comput. Control, Imperial College, London. *Math. Programming Stud.* **16**, (1982), to appear.

Miele, A., Cragg, E. E., Iyer, R. R., and Levy, A. V. (1971a). Use of the augmented penalty function in mathematical programming problems, Part I. *J.O.T.A.* **8**, 115–130.

Miele, A., Cragg, E. E., and Levy, A. V. (1971b). Use of the augmented penalty function in mathematical programming problems, Part II. *J.O.T.A.* **8**, 131–153.

Miele, A., Moseley, P. E., and Cragg, E. E. (1972). On the method of multipliers for mathematical programming problems. *J.O.T.A.* **10**, 1–33.

Mifflin, R. (1977). An algorithm for constrained optimization with semismooth functions. *Math. Oper. Res.* **2**, 191–207.

Moré, J. J., and Sorensen, D. C. (1979). On the use of directions of negative curvature in a modified Newton method. *M.P.* **16**, 1–20.

Mukai, H., and Polak, E. (1975). A quadratically convergent primal-dual algorithm with global convergence properties for solving optimization problems with equality constraints. *M.P.* **9**, 336–349.

Murray. W. (1972). Second derivative methods. *In* "Numerical Methods for Unconstrained Optimization" (W. Murray, ed.), pp. 57–71. Academic Press, New York.

Nguyen, V. H., and Strodiot, J. J. (1979). On the convergence rate of a penalty function method of exponential type. *J.O.T.A.* **27**, 495–508.

Oren, S. S. (1973). Self-scaling variable metric algorithms without line search for unconstrained minimization. *Math. Comput.* **27**, 873–885.

Oren, S. S. (1974). Self-scaling variable metric algorithm, Part II. *M.S.* **20**, 863–874.

Oren, S. S. (1978). A combined variable metric conjugate gradient algorithm for a class of large-scale unconstrained minimization problems. *In* "Optimization Techniques, Part II" (J. Stoer, ed.), pp. 107–115. Springer-Verlag, Berlin and New York.

Oren S. S., and Luenberger, D. G. (1974). Self-scaling variable metric algorithm, Part I. *M.S.* **20**, 845–862.

Oren, S. S., and Spedicato, E. (1976). Optimal conditioning of self-scaling variable metric algorithms. *M.P.* **10**, 70–90.

Ortega, J. M., and Rheinboldt, W. C. (1970). "Iterative Solution of Nonlinear Equations in Several Variables." Academic Press, New York.

Papavassilopoulos, G. (1977). Algorithms for a class of nondifferentiable problems. M.S. Thesis, Dept. Electr. Engr., Univ. of Illinois, Urbana. Also *J.O.T.A.* **34**, (1981), 41–82.

Pietrzykowski, T. (1969). An exact potential method for constrained maxima. *SIAM J. Numer. Anal.* **6**, 269–304.

Polak, E. (1971). "Computational Methods in Optimization: A Unified Approach." Academic Press, New York.

Polak, E. (1976). On the global stabilization of locally convergent algorithms. *Automatica—J. IFAC* **12**, 337–342.

Polak, E., and Ribiere, G. (1969). Note sur la convergence de méthodes de directions conjugées. *Rev. Fr. Inform. Rech. Oper.* **16-R1**, 35–43.

Polak, E., and Tits, A. L. (1979). "A Globally Convergent, Implementable Multiplier Method with Automatic Penalty Limitation," Memo No. UCB/ERL M79/52. Electron. Res. Lab., Univ. of California, Berkeley. *Applied Math. and Optim.* **6**, (1980), 335–360.

Poljak, B. T. (1963). Gradient methods for the minimization of functionals. *Ž. Vyčisl. Mat. i Mat. Fiz.* **3**, 643–653.

Poljak, B. T. (1969a). The conjugate gradient method in extremal problems. *Ž. Vyčisl. Mat. i Mat. Fiz.* **9**, 94–112.

Poljak, B. T. (1969b). Minimization of unsmooth functionals. *Ž. Vyčisl. Mat. i Mat. Fiz.* **9**, 14–29.

Poljak, B. T. (1970). Iterative methods using Lagrange multipliers for solving extremal problems with constraints of the equation type. *Ž. Vyčisl. Mat. i Mat. Fiz.* **10**, 1098–1106.

Poljak, B. T. (1971). The convergence rate of the penalty function method. *Ž. Vyčhisl. Mat. i Mat. Fiz.* **11**, 3–11.

Poljak, B. T. (1979). On Bertsekas' method for minimization of composite functions. *Internat. Symp. Systems Opt. Analysis* (A. Bensoussan and J. L. Lions, eds.). pp. 179–186. Springer-Verlag, Berlin and New York.

Poljak, B. T., and Tretjakov, N. V. (1973). The method of penalty estimates for conditional extremum problems. *Ž. Vyčhisl. Mat. i Mat. Fiz.* **13**, 34–46.

Poljak, B. T., and Tretjakov, N. V. (1974). An iterative method for linear programming and its economic interpretation. *Matecon* **10**, 81–100.

Powell, M. J. D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *Comput. J.* **7**, 155–162.

Powell, M. J. D. (1969). A method for nonlinear constraints in minimization problems. *In* "Optimization" (R. Fletcher, ed.), pp. 283–298. Academic Press, New York.

Powell, M. J. D. (1970). A new algorithm for unconstrained optimization. *In* "Nonlinear Programming" (J. B. Rosen, O. L. Mangasarian, and K. Ritter, eds.), pp. 31–65. Academic Press, New York.

Powell, M. J. D. (1971). On the convergence of the variable metric algorithm. *J. Inst. Math. Appl.* **7**, 21–36.

Powell, M. J. D. (1973). On search directions for minimization algorithms. *M.P.* **4**, 193–201.

Powell, M. J. D. (1977). Restart procedures for the conjugate gradient method. *M.P.* **12**, 241–254.

Powell, M. J. D. (1978a). Algorithms for nonlinear constraints that use Lagrangian functions. *M.P.* **15**, 224–248.

Powell, M. J. D. (1978b). The convergence of variable metric methods for nonlinearly constrained optimization calculations. In "Nonlinear Programming 3" (O. L. Mangasarian, R. Meyer, and S. Robinson, eds.), pp. 27–63. Academic Press, New York.

Pschenichny, B. N. (1970). Algorithms for the general problem of mathematical programming. *Kibernetika (Kiev)*, **6**, 120–125.

Pschenichny, B. N., and Danilin, Y. M. (1975). "Numerical Methods in Extremal Problems." MIR, Moscow. (Engl. transl., 1978.)

Ritter, K. (1973). A superlinearly convergent method for minimization with linear inequality constraints. *M.P.* **4**, 44–71.

Robinson, S. M. (1974). Perturbed Kuhn–Tucker points and rates of convergence for a class of nonlinear programming algorithms. *M.P.* **7**, 1–16.

Rockafellar, R. T. (1970). "Convex Analysis." Princeton Univ. Press, Princeton, New Jersey.

Rockafellar, R. T. (1971). New applications of duality in convex programming. *Proc. Confer. Probab., 4th, Brasov, Romania*, pp. 73–81.

Rockafellar, R. T. (1973a). A dual approach to solving nonlinear programming problems by unconstrained optimization. *M.P.* **5**, 354–373.

Rockafellar, R. T. (1973b). The multiplier method of Hestenes and Powell applied to convex programming. *J.O.T.A.* **12**, 555–562.

Rockafellar, R. T. (1974). Augmented Lagrange multiplier functions and duality in nonconvex programming. *S.J.C.* **12**, 268–285.

Rockafellar, R. T. (1976a). Monotone operators and the proximal point algorithm. *S.J.C.O.* **14**, 877–898.

Rockafellar, R. T. (1976b). Augmented Langrangians and applications of the proximal point algorithm in convex programming. *Math. Oper. Res.* **1**, 97–116.

Rockafellar, R. T. (1976c). Solving a nonlinear programming problem by way of a dual problem. *Symp. Mat.* **27**, 135–160.

Rupp, R. D. (1972). Approximation of the Classical Isoperimetric Problem. *J.O.T.A.* **9**, 251–264.

Sargent, R. W. H., and Sebastian, D. J. (1973). On the convergence of sequential minimization algorithms. *J.O.T.A.* **12**, 567–575.

Schnabel, R. B. (1980). Determining feasibility of a set of nonlinear inequality constraints. *Math. Programming Stud.* **16**, (1982), to appear.

Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Math. Comput.*, **24**, 647–656.

Shanno, D. F. (1978a). Conjugate gradient methods with inexact searches. *Math. Oper. Res.* **3**, 244–256.

Shanno, D. F. (1978b). On the convergence of a new conjugate gradient algorithm. *SIAM J. Numer. Anal.* **15**, 1247–1257.

Shapiro J. E. (1979). "Mathematical Programming Structures and Algorithms." Wiley, New York.

Shor, N. Z. (1964). On the structure of algorithms for the numerical solution of planning and design problems. Thesis. Kiev.

Shor, N. Z. (1970). Utilization of the operation of space dilation in the minimization of convex functions. *Kiberbetika (Kiev)* **6**, 6–12.

Shor, N. Z., and Jourbenko, N. G. (1971). A method of minimization using space dilation in the direction of two successive gradients. *Kibernetika (Kiev)* **7**, 51–59.

Sorenson, H. W. (1969). Comparison of some conjugate direction procedures for functions minimization. *J. Franklin Inst.* **288**, 421–441.

Stephanopoulos, G., and Westerberg, A. W. (1975). The use of Hestenes' method of multipliers to resolve dual gaps in engineering system optimization. *J.O.T.A.* **15**, 285–309.

Stoilow, E. (1977). The augmented Lagrangian method in two-level static optimization. *Arch. Automat. Telemech.* **22**, 219–237.

Tapia, R. A. (1977). Diagonalized multiplier methods and quasi-Newton methods for constrained minimization. *J.O.T.A.* **22**, 135–194.

Tapia, R. A. (1978). Quasi-Newton methods for equality constrained optimization: Equivalence of existing methods and a new implementation. *In* "Nonlinear Programming 3" (O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds.), pp. 125–164. Academic Press, New York.

Vastola, K. S. (1979). A numerical study of two measures of delay for network routing. M.S. Thesis. Dept. Electr. Engr., Univ. of Illinois, Champaign-Urbana.

Watanabe, N., Nishimura, Y., and Matsubara, M. (1978). Decomposition in large system optimization using the method of multipliers. *J.O.T.A.* **25**, 181–193.

Wierzbicki, A. P. (1971). A penalty function shifting method in constrained static optimization and its convergence properties. *Arch. Automat. Telemech.*, **16**, 395–416.

Wilson, R. B. (1963). A simplicial algorithm for concave programming. Ph.D. Thesis, Grad. Sch. Business Admin., Harvard Univ., Cambridge, Massachusetts.

Wolfe, P. (1969). Convergence conditions for ascent methods. *SIAM Rev.*, **11**, 226–235.

Wolfe, P. (1975). A method of conjugate subgradients for minimizing nondifferentiable functions. *Math. Programming Stud.* **3**, 145–173.

Zangwill, W. I. (1967a). Minimizing a function without calculating derivatives. *Comput. J.* **10**, 293–296.

Zangwill, W. I. (1967b). Nonlinear programming via penalty functions. *M.S.* **13**, 344–358.

Zangwill, W. I. (1969). "Nonlinear Programming." Prentice-Hall, Englewood Cliffs. New Jersey.